NBER WORKING PAPER SERIES

CAN YOU RECOGNIZE AN EFFECTIVE TEACHER WHEN YOU RECRUIT ONE?

Jonah E. Rockoff
Brian A. Jacob
Thomas J. Kane
Douglas O. Staiger

Working Paper 14485
http://www.nber.org/papers/w14485

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2008

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. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Can You Recognize an Effective Teacher When You Recruit One?
Jonah E. Rockoff, Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger
NBER Working Paper No. 14485
November 2008
JEL No. I21,J45

ABSTRACT

Research on the relationship between teachers' 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. However, most research has examined a relatively small set of characteristics that are collected by school administrators in order to satisfy legal requirements and set salaries. 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 non-traditional 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. Individually, we find that only a few of these predictors have statistically significant relationships with student and teacher outcomes. However, when all of these variables are combined into two primary factors summarizing cognitive and non-cognitive teacher skills, we find that both factors have a modest and statistically significant relationship with student and teacher outcomes, particularly with student test scores. These results suggest that, while there may be no single factor that can predict success in teaching, using a broad set of measures can help schools improve the quality of their teachers.

Jonah E. Rockoff
Columbia University
Graduate School of Business
3022 Broadway #603
New York, NY 10027-6903
and NBER
jonah.rockoff@columbia.edu

Thomas J. Kane
Harvard Graduate School of Education
Gutman Library, Room 455
Appian Way
Cambridge, MA 02138
and NBER
kaneto@gse.harvard.edu

Brian A. Jacob
Gerald R. Ford School of Public Policy
University of Michigan
735 South State Street
Ann Arbor, MI 48109
and NBER
bajacob@umich.edu

Douglas O. Staiger
Dartmouth College
Department of Economics
HB6106, 301 Rockefeller Hall
Hanover, NH 03755-3514
and NBER
douglas.staiger@dartmouth.edu

"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 teachers' 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 et al. (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 et al. (2005), Kane et al. (2006)), and Aaronson et al. (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 et al. (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 2006-2007. 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 non-traditional 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 the district and school level. In addition to comparing the predictive power of our non-traditional 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 widely used to screen candidates—the Haberman Star Teacher Evaluation PreScreener. The Haberman PreScreener is used by a number of large urban school districts throughout the U.S., and is intended to provide school officials with guidance on how effective a particular candidate is likely to be in an urban classroom. 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 statistically significant but modest relationships between student achievement and several non-traditional predictors of teacher effectiveness, including performance on the Haberman selection instrument. We find marginally significant increases of about 0.02 standard deviations in math achievement associated with one-standard deviation increases in cognitive ability and self-efficacy. For respondents' scores on a test of math knowledge for teaching, we

estimate an effect size of about 0.03 standard deviations and statistical significance at the 2 percent level. Scores on the Haberman PreScreener are also positively related to student achievement, with an effect size of 0.02 standard deviations, which is marginally significant at the 11 percent level. As a point of comparison, prior research using similar data from the DOE (Kane et al. (2006)) found that assignment to a teacher with a year of teaching experience in the DOE, as opposed to a true "rookie" is associated with roughly 0.04 standard deviations higher student achievement. Interestingly, we do not find 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.

No single metric we examine has the ability to reliably identify very large differences in teacher effectiveness among our survey respondents. However, through the use of factor analysis, we document how these metrics can be combined into simpler measures of cognitive and non-cognitive 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 one standard deviation higher on either the cognitive or non-cognitive factor have achievement that is .033 standard deviations higher. These results suggest that schools and school districts wishing to increase the effectiveness of their teacher workforce may be aided by the systematic use of a broad set of information on new candidates, and particularly if they gather information outside the realm of traditional teaching credentials. Nevertheless, our results

are also consistent with the notion that data on job performance may be a more powerful tool for improving teacher selection than data available at the recruitment stage.

The paper proceeds as follows. In Section 2 we describe the contents of our survey of new teachers and in Section 3 we provide details on our sample and the additional data we use to examine student and teacher outcomes. Section 4 provides descriptive statistics on survey respondents and their responses. We present our methodology and the results of our analysis of traditional and non-traditional predictors in Section 5. Section 6 presents results of a factor analysis on teacher characteristics and tests of the predictive power of these factors for student and teacher outcomes. Section 7 concludes.

2. Survey Elements

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 2006-2007. 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. In this section, we provide details on each of the major survey components, describing the theory and research that motivated its inclusion in the survey. We provide examples of many of the items in the appendix, and the full survey is available upon request. 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 reviews of this literature, see Jacob (2007).

2.1 A Teacher's Cognitive Ability and Academic Success

Some researchers have found that teachers with stronger academic backgrounds produce larger performance gains for their children (see, for example, Clotfelter et al. (2006, 2007), in

addition to the reviews cited above). However, there are also a number of studies which do not find this relationship (e.g., Harris and Sass (2006) on graduate course work and Kane et al. (2006) on college selectivity). In our survey, we collect a number of measures of academic success, covering many of the measures used by prior researchers (e.g., undergraduate major, graduate education, selectivity of undergraduate institution, etc.).[1]

A small number of studies have found a link between teachers' scores on certification examinations and teacher effectiveness (e.g., Clotfelter et al. (2006, 2007) and Goldhaber (2007), although Harris and Sass (2006) do not find this link). Although teachers in New York State must take several exams in order to become legally certified to teach, the New York City DOE does not have access to teacher certification exam scores, and these scores are unlikely to be known by district personnel making hiring decisions. According to New York State, these exams "are for the purpose of New York State educator certification only. They are not intended to be used for employment decisions, college admissions screening, or any other purpose. Candidates are not obligated to provide potential employers with copies of [their] score reports."[2]

The main certification test in New York State is the Liberal Arts and Science Test (LAST), which is required for certification in all subjects and can be taken an unlimited number of times until a teacher passes. Boyd et al. (2006, 2008a, 2008b) use whether a teacher that passed the LAST on the first attempt as a marker of effectiveness, and find mixed results. We

[1] We asked respondents for their undergraduate institution, and we merge this information to the Barron's Selectivity Index (a 1-9 scale, one being the best) from 1982. We thank Caroline Hoxby for sharing this data with us. For a few colleges where the Barron's rating was missing, we use Barron's ratings from 1984.
[2] See www.nystce.nesinc.com and ohe32.nysed.gov/tcert/ for general information on certification exams and www.nystce.nesinc.com/pdfs/NYSTCE_ISR_back.pdf for information on the use of exam scores.

therefore asked teachers whether they passed the LAST on their first attempt and examine this variable below.[3]

While several early studies failed to find a significant relationship between college admissions scores and principals' evaluations of new teachers (e.g., Maguire (1966), Ducharme (1970)), a commonly cited study by Ferguson and Ladd (1996) did find a link between scores on the ACT exam and student achievement growth. We therefore asked teachers about their college entrance examination scores. While we asked specifically about both the SAT and the ACT, few teachers reported an ACT score and, of those that did, 90 percent also reported an SAT score. We therefore do not use the ACT in our analysis. Although nearly 80 percent of the respondents claimed to have taken the SAT, less than one in three reported their exact scores. Anticipating that some teachers might not remember their scores, we also allowed teachers to give their scores in 100 point ranges, which most did, and we assign these teachers the midpoint of the reported range (e.g., we assign a score of 550 for someone reporting a score between 500 and 600). Still, about 50 teachers (12 percent of respondents) reported that they took the SAT but could not remember their scores at all.

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 resources in childhood, parental investment in education, personal motivation and willingness to study hard, raw intelligence, etc. In order to separate out at least one of these proximate causes, the survey includes a direct test of cognitive

[3] 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.

ability, Raven's Progressive Matrices Standard Version, an intelligence test that requires no linguistic or mathematics skills.[4] An illustrative item for this instrument (taken from Raven (2000)) is shown in Appendix Figure 1. We convert scores on this cognitive ability test to national percentiles using the distribution for a representative sample of U.S. adults ages 20-47 who completed the self-administered test at leisure (Raven et al., 2000).

2.2 Content Knowledge

A number of studies examine the relationship between content knowledge and effectiveness, particularly in teaching mathematics (e.g., Goldhaber and Brewer (1997), Aaronson et al. (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 which 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 (Schulman (1986, 1987), Wilson et al. (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 mathematical knowledge among teachers (Hill (2006)). There is evidence of a positive relationship between content knowledge (as measured by this instrument) and student achievement gains in first and third grade (Hill et al. (2005)). Importantly, they also found this

[4] 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. 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 non-verbal nature. The split-half reliabilities for this test are also high, with a coefficient of .86 (Raven et al. (1983)).

measure to be a stronger predictor of student learning than other measures of teachers' mathematical preparation. An item from this instrument is presented in Appendix Figure 2.

2.3 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 much of this work focuses on comparing attitudes across teachers and other occupations, or across specialties within teachers, a few studies (e.g., Washburne and Heil (1960)) linked child-friendly attitudes with positive teaching outcomes (although no studies assess student achievement directly). While many studies have been conducted, few definitive conclusions have been made. One reason has been the widespread but controversial use of the Minnesota Multiphasic Personality Inventory (MMPI) to measure teacher personality traits, even though the MMPI was designed to measure social and behavioral problems in psychiatric patients. Getzels and Jackson (1963) find no consistent relationship between personality traits as measured by the MMPI and measures of teacher success. Another reason why clear predictions have been difficult in this field is the wide variety of theories and measures of personality that abound in psychology. However, recent decades have seen a move from theorist-driven accounts of personality (dominated by Freud and Jung) to simple empirical measures of important dimensions of personality.

One such empirical model, the five-factor model (or "Big Five"), has emerged as a dominant new framework for measuring personality. The Big Five personality traits are: agreeableness, conscientiousness, emotional stability, extraversion, and openness to experience. We are not aware of any work linking elements of the Big Five to teacher effectiveness in raising student achievement. However, the Big Five have been used to predict job performance across a wide variety of other occupations. Using meta-analysis, Barrick and Mount (1991) find that

conscientiousness has been linked positively to job performance across all occupational categories. They also document a link between extraversion and job performance in occupations requiring social interaction. Similar results are echoed in a review by Goodstein and Lanyon (1999). Thus, we hypothesize that conscientiousness and extraversion may be significant predictors of job performance for teachers.

Instruments used to measure the Big Five vary in length and complexity. We employ the Big Five Inventory (BFI), developed by John et al. (1991), which consists of 44 items: 10 for openness to new experience, 9 each for agreeableness and conscientiousness, and 8 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.

2.4 Teacher Beliefs and Values

The idea of self-efficacy—the belief that one can successfully produce an outcome—as an important factor in determining whether individuals can overcome challenges and meet goals is well established in the field of psychology (see Bandura (1977)). Moreover, 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 et al. (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. One exception is an oft-

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 paper also finds a significant positive relationship between teachers' sense of self-efficacy and student achievement growth.[5]

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 impact student learning) and general efficacy (i.e., belief in the ability of teachers in general to impact 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.

2.5 Teacher Selection Instruments

One ultimate policy goal of research on predictors of teacher effectiveness is to develop tools which district and school administrators could use to identify the "most promising" teacher candidates. However, there are already two commercially available and widely used instruments whose purpose is to measure beliefs and values indicative of future success in the classroom: the Haberman Star Teacher Evaluation PreScreener ("Haberman PreScreener") and the Gallup TeacherInsight Assessment (Gallup TIA). The two instruments are similar in that they both use a short survey consistent mostly of multiple choice items to evaluate a number of teachers'

[5] The two questions used by Armor et al. (1976) to measure efficacy are included in our measures—one as part of the general efficacy index and one as part of the personal efficacy index. Notably, their study, like ours, uses data on teachers' self-efficacy collected after the start of the teachers' careers.

attributes.[6] Both the Haberman PreScreener and the Gallup TIA were developed by first interviewing teachers thought to be highly effective and designing questions to capture their attitudes and beliefs. These instruments have been used by many large urban school districts throughout the U.S., 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. Haberman (1993, 1995) has published some reports of his research, but no empirical data are available for independent analysis. New York City recently began requiring all applicants for teaching positions to take the TIA. In ongoing work, we are assessing how well this instrument predicts student and teacher outcomes in the district. In this paper, we analyze the Haberman PreScreener, which was included as a part of our survey and was scored for us by the Haberman Foundation. Each teacher is given a categorical score of "Low," "Average," or "High" in each of ten attributes (see footnote 9) and an overall score for the total number of questions answered correctly. In their work with districts, the Haberman Foundation places teacher candidates into four ranked categories: 1) a top group which includes candidates who answered at least 33 questions correctly, and did not receive a "low" score in any of the ten categories; 2) a second group which includes candidates who did not receive any "low" scores but answered less than 33 questions correctly; 3) a third group which includes candidates who answered at least 33

[6] The Haberman PreScreener is a short survey that uses 50 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. Similarly, the TIA instrument uses multiple choice, Likert scale (i.e., level of agreement from 1 to 5), and open-ended items to assess a number of teacher attributes. We have been unable to find a list of attributes for the Gallup TIA, but an earlier Gallup instrument, the Teacher Perceiver Interview, measured 12 attributes (Metzger and Wu, forthcoming): Mission, Empathy, Rapport drive, Individualized perception, Listening, Investment, Input drive, Activation, Innovation, Gestalt, Objectivity, and Focus.

questions correctly, but had a "low" score in one of the ten categories; and 4) a bottom group that consists of teachers who either (i) received one low score and answered less than 33 questions correctly or (ii) received two or more low scores regardless of the number of questions answered correctly. According to Haberman officials, no applicant with two or more "low" scores should be hired, regardless of the total number of questions correct.[7]

Twenty-one percent of our survey respondents completing the Haberman PreScreener fell into the top group according to the categorization system described above, while 60 percent fell into the bottom group. In our analysis, we test whether being in the top group of teachers is predictive of positive outcomes. However, we make use of the other variation in the data by testing the predictive power of the total number of questions answered correctly.[8]

2.6 Other Teacher Characteristics

In addition to the items described above, we also asked about several other characteristics. These included occupations prior to teaching in the DOE, weeks and hours per week of paid and volunteer experience in various fields related to working with children (i.e., full-time teaching, substitute teaching, work as an education paraprofessional, tutoring, work in after-school programs, coaching, baby-sitting, work in child care/day care, camp counselor, work in community programs, mentor, and work in religious education), childhood setting (i.e., rural, suburban, urban, or foreign), K-12 education (public or private), and attendance of New York City public schools. In preliminary analyses not reported here (but available upon request), we found no systematic and/or significant relationship between these measures and our outcomes.

[7] Description of the 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.
[8] Note that this is not based on any recommendation of Martin Haberman or the Haberman Foundation.

3. Data Collection and Analysis Sample

Here we describe more carefully the administration of the survey, the administrative data used to measure student and teacher outcomes, and the construction of our analysis sample.

3.1 Survey Administration

Due to budget constraints, we target our survey to new elementary and middle school math teachers, a group for whom we could calculate value-added measures of effectiveness using models that relied on prior test scores as a control. With the assistance of DOE officials, we identified 602 new teachers with no prior experience who were identified as teaching mathematics to students in grades four through eight (testing begins in third grade in New York City). Some of these teachers were teaching all subjects to a single elementary class, while others taught math to one or more classrooms of students in middle school grades.[9]

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 April 3, 2007, and teachers were given until the end of June to complete the survey.[10] 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 200 items. Pilot testing of the survey with students at the Harvard Graduate School of Education suggested that completion

[9] In general, elementary schools in New York City include grades K-5, middle schools include grades 6-8 and high schools include grades 9-12. However, there are schools with a variety of different grade configurations, such as K-8, 5-8, 6-7, 6-12, etc.

[10] In order to protect the confidentiality of the data, communication with teachers was done via the Human Resources Department at the DOE. 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.

would require about 90 minutes. 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 non-respondents and non-completers between the start 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.[11] In Section 4, we compare respondents and non-respondents on a variety of observable characteristics.

3.2 Administrative Data

In addition to the responses to our survey, we use data from a number of other sources in our analysis. Administrative data from the 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. This provides information on 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 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 four through eight. These data follow students over time and provide links to their math teachers. The student data we possess also 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 et al. (2006).

A small but growing literature demonstrates a significant relationship between objective measures of teacher performance and subjective evaluations of teacher quality made during a

[11] Respondents include all teachers who began the survey, including 15 teachers who began the survey but did not complete any of the main sections. Placing these 15 teachers in the non-respondent category does not noticeably our comparisons of respondents and non-respondents (Table 1).

teacher's career (e.g., Murnane (1975), Armor et al. (1979), Harris and Sass (2008), and Jacob and Lefgren (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. These data come from a centrally administered program to assist new teachers, which was created to comply with a New York State law requiring mentoring (see Rockoff (2008)). We do not have evaluations for new teachers in a number of schools that were exempt from the centralized mentoring program due to their status as an "Empowerment School," which gave more programmatic choice to principals.[12]

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 towards the end of the year. 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 over ratings given in the months of April, May, and June. For the teachers who were not rated in those months (less than two percent of teachers with any recorded evaluations), we use ratings averaged over January, February, and March.

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' Common Core of Data. This includes school level data on student ethnicity, gender, and eligibility for free

[12] For more information on Empowerment schools, see http://schools.nyc.gov/Offices/Empowerment/ .

lunch of students, as well as the school's eligibility for Title I resources, pupil-teacher ratio, and grade composition. In order to better control for differences across schools that are unobservable in the CCD data but 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.

3.3 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 (subjective evaluations, absences, and retention) we include data on the 184 teachers who were asked to take the survey but did not respond and a set of 4,275 other new teachers. This set of other new teachers are defined as those with no prior teaching experience that started in the school year 2006-2007 who were present in the DOE payroll files in both November and May, did not teach in a special program (e.g., extended high school for adults), were linked with school level data on student characteristics, and were not asked to take our survey.[13] For each of the outcomes that we explore, our sample naturally includes only those teachers with valid outcome data. We have attrition data for all 4,877 teachers in our sample, but lack absence data for 19 teachers. For mentor ratings, we have data on 3,030 teachers (62 percent of our sample). The fraction of teachers with mentor evaluations is somewhat higher among teachers who responded to our survey (75 percent) or were asked to take our survey but did not respond (73 percent) than among those who were not asked (60 percent). Nearly 70 percent of the missing evaluations are due to teachers working in Empowerment schools, which did not participate in

[13] Conditioning on presence in November and May ensures that, like the teachers invited to the survey, the other new teachers were hired close to the start of the school year and did not leave before the end of the year. While conditioning on presence in payroll in September and May might seem more appropriate, the timing of record updating in the DOE is such that many new hires are not present in the September payroll data.

the centralized mentoring program. Of the remaining teachers, 83 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.[14]

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 2006-2007. We include these additional classrooms in order to gain better estimates of the coefficients on important control variables, such as prior student achievement, participation in English Language Learner and special education programs, etc. In addition, we restrict our analysis using the same rules as in Kane et al. (2006): excluding schools where we could not successfully merge at least 75 percent of the classes with teachers and schools serving only special education students (176 out of 1169 schools), classrooms that could not be linked to a teacher (less than 2 percent of classrooms in the remaining sample), where more than 25 percent of students received special education services (19 percent of classrooms in the remaining sample, 73 percent of which had only special education students), which had at least 7 and no more than 45 students (eliminating 10 percent of the remaining classrooms), and whose assigned teacher left mid-year or switched schools (2 percent of remaining classrooms). This leaves us with just over 13,000 classrooms in 988 schools. In total, we are unable to examine math value-added for 43 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, 5 switched schools during the year, and 36 taught in classrooms where more than 25 percent of the students were classified as receiving special education services.

[14] 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 were asked to take our survey but did not respond (92 percent) than among those who were not asked (82 percent).

4. Descriptive Statistics

Table 1 provides summary statistics broken down into three groups: survey respondents, new teachers who were invited and did not respond, and other new teachers hired in 2006-2007 that were not invited to participate in the survey. The third column provides P-values on tests of whether there is a statistically significant difference in the mean of a characteristic between respondents and non-respondents. Of the 18 teacher and school characteristics listed in the table, there are two on which the respondents and non-respondents are significantly different at the five percent level or lower. Relative to non-respondents, respondents were more likely to be female (78 percent vs. 66 percent), and were less likely to come from the Teach for America 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). While we do not report statistical tests of differences between teachers not invited to take our survey and those that were, they are fairly similar along characteristics to the teachers who were invited to take the survey.[15]

Summary statistics on outcomes for all three groups are shown at the top of Table 1. Absences for new teachers averaged 5.7 over the school year for teachers asked to take our survey and 6.4 for those who were not asked. The standard deviation of absences among all teachers in our sample is 4.7, but the distribution is skewed, ranging from 0 to 41. Among survey respondents, 8.1 percent did not return to teaching in the DOE the following school year, similar to 6.5 percent for non-respondents and 7.4 percent for other new teachers. An additional

[15] Though not shown in Table 1, far more teachers invited to take the survey were licensed in math, but this is not surprising given that we targeted our survey to math teachers. We have also compared the characteristics of teachers who completed to the survey to 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 non-White and less likely to come from the Teach for America program.

8.9 percent of respondents returned to teach in a different school within the DOE, compared with 8.2 percent of non-respondents and 8.1 percent of other new teachers.

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.[16] However, there is considerable variation in college major between teachers assigned to students in grades four and five (28 percent majoring in education and 3 percent in math and science) and those assigned to grades six to eight (9 percent majoring in education and 34 percent in math and science). Thirty-two percent of survey respondents reported having a graduate degree. Average reported SAT scores were roughly 600 in both math and verbal with a standard deviation of about 100 points. The fairly high averages may reflect the percentage of Teaching Fellows and TFA corps members in our sample, and perhaps non-random selection in teachers' willingness to report their scores. The average Barron's rank of respondents' undergraduate institutions was 5.6 (on a 1-9 scale with 1 being the highest). Twelve percent of respondents' institutions ranked in the top three categories, with 40 percent in the middle (ranked four to six) and the remainder from institutions ranked seven or below. Nearly all of the respondents (92.2 percent) claimed to have passed the LAST exam on their first attempt. This is somewhat higher than the pass rates for new teachers in the school year 2004-2005, which were less than 90 percent (Boyd et al.

[16] We group all other college majors together in our analysis. About 30 percent of survey respondents majored in political or social sciences, 13 percent in English or humanities, 9 percent in Foreign languages or communications, 7 percent in business, 5 percent in the Arts, and two percent in "Other" (i.e., they did not find a match among the 50 majors we presented as choices).

2006), but may simply reflect a continued trend of increasing pass rates for new teachers in New York City.

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, 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 of these teachers must have a college degree, this is not terribly surprising.

The portion of answers answered correctly 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. 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. 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). Interestingly, while math or science majors scored significantly higher than education majors (60 percent vs. 49 percent correct), respondents with majors other than education, math and science performed similarly well (60 percent correct).[17]

[17] As an additional check on the academic background survey results, we compared scores on cognitive ability, math content, and (self-reported) college entrance examinations for groups of teachers from different certification pathways. On all tests, scores for teachers from the New York City Teaching Fellows program were higher than regularly certified teachers, and scores for teachers from the Teach for America program were higher than both other groups. This matched our expectations; both TFA and the Teaching Fellows recruit candidates from highly selective colleges and universities, but the TFA program is generally recognized as more selective.

In Table 2 we report the raw scores (on a scale of 1-5) for all five dimensions of personality from the Big Five Inventory, though in our analysis below we restrict our attention to conscientiousness and extraversion. While these summary statistics are difficult to interpret, to our knowledge, there is no standard benchmark for the Big Five. The National Survey of Midlife Development in the United States, 1995-1996, did collect data on the Big Five for a representative sample of English-speaking, non-institutionalized, U.S. adults between the ages of 25 and 74.[18] However, the two sets of results are not directly comparable because the exact number and wording of the items in this survey were not identical to ours and because responses were given on a scale of 1 to 4 (see Lachman and Weaver (1997)). Therefore, rather than ask whether survey respondents score higher or lower than the 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 the 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.[19] However, there are no striking differences between the two samples' scores.

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 in 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. However, the variation in scores within all three groups is of similar magnitude. The correlation between personal and

[18] This data is available from ICPSR as Study No. 2760.

[19] 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.

general efficacy 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 average total number of items answered correctly (out of 50) was about 32, with a standard deviation of about five points. Haberman cites 32 as a median score, so that our sample of teachers (for whom the mean and median are both 32) seems to have scored similarly to the population of individuals in other districts that have completed the Haberman instrument.

5. Predictors of Teacher and Student Outcomes

In this section, we examine how well a series of traditional and non-traditional teacher characteristics predict student and teacher outcomes. In Section 5.1, we outline the statistical methodology we use, highlighting some of the limitations of our approach. In Section 5.2, we present results that present each predictor separately in order to measure the overall relationship of each predictor with teacher and student outcomes. In Section 5.3, we investigate the correlates of performance on the Haberman PreScreener and the power of this instrument to predict teacher and student outcomes.

5.1 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_j 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_j is an idiosyncratic error term.

$$(1) \qquad Y_j = \alpha + \delta P_j + \beta X_j + \gamma SC_{jk} + \varepsilon_j$$

As mentioned earlier, we include in our analysis a large number of new teachers who were not asked to take our survey. For these teachers, and for teachers who did not respond to the survey invitation or did not complete a particular part, we set the predictor variable to zero and include an indicator for whether an actual survey response was missing. We do this in order to obtain better estimates of the coefficients on our school-level controls. To the extent that factors such as school poverty (i.e., the fraction of students eligible for free lunch) influences outcomes such as teacher absences, the exclusion of these controls (or mis-measurement of the true effect of these characteristics) may lead to biased estimates of our key predictors.

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.

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

Following the approach described above, we include students taught by teachers who were not invited to take the survey or did not respond in order to identify the coefficients on student and school characteristics. As with teacher outcomes, we use indicators for teachers with missing survey data and set predictor variables to zero for the students assigned to these teachers.

We examine five dependent variables in our analysis: student test scores in math, teacher absences, subjective evaluations of teachers, 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 one so that coefficients can be readily interpreted. In order to maximize our statistical power in examining predictors from our survey, we include all individuals with non-missing data, so that, while our

sample size does not vary across the specifications, the true number of teachers with identifying variation fluctuates slightly. For simplicity in exposition, we use linear regression analysis in all cases, and report coefficients and standard errors clustered at the school level. We find very similar results to those presented here using negative binomial regressions to examine absences and conditional logistic regression to examine teacher retention.

In all regressions, we include controls for the characteristics of schools (from the Common Core of Data), school zip code fixed effects, and grade level fixed effects. In the student achievement specifications, we drop the school average characteristics from the CCD but include controls for individual students' prior student test scores (specifically, cubic polynomials in both prior math and reading scores, interacted with grade level), 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), as well as classroom and school averages of these student characteristics. We regard this specification as generating valid estimates of the relationship between survey variables and teacher effectiveness. While we recognized 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.

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, for example, that

extraversion and empathy are positively correlated and both positively impact student achievement. In this case, the exclusion of empathy from our estimates may lead us to overstate the effect of extraversion on student performance.

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, then 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.[20]

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

[20] In addition, if one knew the true "structural" relationship between teacher characteristics and effectiveness, then one might develop professional development to enhance those characteristics that lead to effectiveness.

school administrators will bias our results towards zero. However, this type of bias only occurs 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 our survey until April 2007. One might be concerned that some of our estimates reflect reverse 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) influence how individuals respond to the Haberman instrument, one should be cautious about interpreting the coefficients on this measure as well.

5.2 The Power of Individual Predictors of Teacher Effectiveness

Table 3 shows results for the power of traditional credentials for predicting each of our five outcomes measures. Within each column, dotted lines separate coefficient estimates from regressions in which we include a single predictor or group of related predictors. The first column presents results for student achievement in math, our primary outcome of interest. Consistent with many other researchers, we find no significant relationship between graduate education and teacher effectiveness; indeed, the coefficient is negative. We do not find that

respondents who passed the main state certification "basic skills" exam – the Liberal Arts and Science Test (LAST) - on the first attempt are significantly more effective, but it is worth noting that very few survey respondents (8 percent) reported failing this exam. 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).

When comparing alternatively certified teachers to traditionally certified among the survey respondents, we find that teaching fellows are less effective (-0.05 standard deviations, p-value = 0.09) and Teach for America corps members are more effective (0.04 standard deviations (p-value = 0.15).[21] While the result on TFA is consistent with other findings (Decker et al. (2004), Boyd et al. (2006), Kane et al. (2006)), the negative finding for teaching fellows contrasts with earlier work (Boyd et al. (2006), Kane et al. (2006)). Non-random 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 non-fellows) hired in earlier years. Thus, it appears to be the case that either this particular group of Teaching Fellows is relatively less effective than earlier cohorts, or that the gains to experience for Teaching Fellows are greater than for other teachers. Although we cannot distinguish these two explanations without additional data, Kane et al. (2006) present some evidence in support of the latter hypothesis.

Students' test scores growth was greater on average with respondents who majored in math or science (0.04 standard deviations, p-value = 0.2) and slightly lower with respondents

[21] While we include controls for other alternative route programs (e.g., the Peace Corps Fellows) there are far fewer teachers in these programs and only a handful in our survey sample, and we do not report their coefficients.

who majored in education (-0.009, p-value = 0.79); we cannot reject that the coefficients are equal (p-value = 0.24). Respondents' self-reported SAT math and verbal scores are also not significantly related to teacher effectiveness. However, the selectivity of respondents' undergraduate institutions, as measured by the Barron's scale, is positive and marginally significant (p-value = 0.08). The positive, albeit small, relationship between college selectivity and teacher effectiveness has been found in other studies (e.g., Clotfelter et al. (2007), Boyd et al. (2008a)). The lack of statistical significance for SAT scores contrasts with findings from other research, but it is worth pointing out again that these scores are self-reported and often reported in ranges, so that measurement error (both classical and systematic) may be pushing the coefficient estimates towards zero.

Turning to the teacher level outcomes in Table 3, the only traditional credential that is related to subjective evaluations is college selectivity, with 0.2 standard deviation lower evaluations given to respondents that attended a college with a ranking one standard deviation above average. We find no statistically significant difference in the average evaluation given to respondents that were alternatively certified vs. traditionally certified. We do, however, find that teaching fellows were absent approximately 1 day more on average than other respondents, and that math and science majors were absent about 1.2 days less. No other traditional credentials were significant predictors of absences.

With regard to retention, we find negative effects for having a graduate degree (-0.05, p-value = 0.13) and being an education major (-.10, p-value = 0.02) 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

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). First-year retention rates for TFA corps members, before their commitment has ended, are typically quite high, but retention after the second year is markedly lower (see Kane et al. (2006)). Conditional on returning to teach in the DOE, TFA corps members are also more likely to return to the same school. This may, however, be driven by the fact that TFA works directly with a limited number of schools to fill positions in high needs areas.

Table 4 presents results on the predictive power of the non-traditional measures gathered in our survey. All of these measures have been normalized, so that the coefficients can be interpreted as the estimated effect of moving one standard deviation in the distribution of the predictor. Again, within each column, dotted lines separate coefficient estimates from regressions in which we include a single predictor or group of related predictors. As above, note that 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 marginally significant (p-value = 0.17) with a coefficient of 0.016, suggesting that cognitive ability may bear some relation to teacher effectiveness. Math knowledge for teaching is more strongly related to math achievement, with a coefficient of 0.028 which is statistically significant at the 2 percent level. This gives support to the work by Hill et al. (2005), who found this instrument to be a significant predictor of teacher effectiveness and a better predictor than other measures of teachers' math training.

The coefficients on conscientiousness (0.011) and extraversion (0.007) are positive, but not significant at conventional levels (p-values of 0.32 and 0.52, respectively). For general and personal efficacy, we also find positive coefficients (0.017 and 0.012, respectively) with marginal significance on general efficacy (p-value = 0.15). Overall, these results give mild support to the idea that teachers' personalities and attitudes are related to teacher effectiveness.[22]

Interestingly, when we consider the relationship between these non-traditional measures and the subjective evaluations of teachers provided by mentors, we find very different results. Subjective evaluations are significantly higher for respondents with high levels of conscientiousness, extraversion, and high levels of personal efficacy, and the coefficients are quite large, ranging from 0.19 to 0.22. In contrast, the evaluations bear little relation to the three non-traditional variables that were (marginally) significant predictors of math achievement, though these coefficients are positive.

Given the 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 one-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

[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 or not a teacher focuses on the test as an important outcome. However, the point estimates on both of 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 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] Author's calculations are available upon request. The use of these subjective evaluations by mentors as a means for identifying effective teachers after the recruitment stage is the subject of ongoing research by one of the authors.

evaluations is clearly due to factors unrelated to 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. Most of the studies in this literature use subjective evaluations of employee performance by supervisors as the outcome of interest. Our findings here suggest that subjective evaluations may be driven by both worker productivity and other worker characteristics, but that some worker characteristics that correlate with evaluations may be unrelated to productivity.

With regard to absences, respondents with cognitive ability scores or math knowledge for teaching scores one standard deviation above average were absent 0.4 days less.[24] Respondents with general efficacy scores one standard deviation above average were more likely to return to the DOE. 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. However, it is worth noting that the questions regarding personal efficacy, as opposed to general efficacy, are more focused on the teacher's own ability to succeed in the classroom, yet the retention result shows up for general efficacy, as opposed to personal.

Overall, the results presented in Tables 3 and 4 suggest that both traditional and non-traditional predictors may be associated with teacher 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 coefficients being considered, any reasonable adjustment for testing multiple hypotheses would make these associations appear even less significant. Second, even the fact that many of the

[24] Because the distribution of absences is skewed, we also examined the natural log of absences and an indicator for having 8 or more absences (corresponding to the 75th percentile or higher) and found similar qualitative results.

coefficients are in the expected direction may simply reflect the fact that many of the predictors are 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 et al. (2006) 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 (.028 on math knowledge for teaching) implies that we are predicting less than 8% of the teacher-level variation.

5.3 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. We first explore what characteristics and traits the Haberman PreScreener captures, and then determine how well it predicts student and teacher outcomes.

Unlike the other non-traditional measures in our survey, the Haberman PreScreener is designed to evaluate a number of characteristics of teachers 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 non-traditional measures of teacher effectiveness included in Tables 3 and 4. Our dependent variables are whether the respondent placed in the "top group" using Haberman's method of screening candidates (i.e., a total score above 32 and zero "low" scores in any of ten categories) and the respondent's total score. We present results that include each measure as a single

predictor in separate regressions that also control for grade level taught and the school 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 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 non-traditional 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.[25] Thus, as we expected, the questions on the Haberman Pre-screener 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 here as we used 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: being in the top group of candidates and total score. While we do not find that being in the top group of candidates is significantly related to our outcome variables, we do find stronger relationships when examining respondents' total scores. A one standard deviation increase in the score on the Haberman PreScreener is associated with a 0.023 standard

[25] 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 lock step. 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 50 questions focus 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.

deviation increase in math achievement that is marginally significant (p-value 0.11) and a 0.14 standard deviation increase in subjective evaluation (p-value = 0.03). Increases in the score were also associated with a greater propensity to return to teaching the following year, although they also predicted a higher probability of transferring to another school within the DOE conditional on returning to teach.[26] While these results should be taken 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.

6. 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, then 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 of 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 of the variables whose coefficients are shown in Tables 3 and 4, as well as the Haberman total score. We do not include the indicator for being in the top group according to Haberman scoring methodology; the total score has a stronger relationship with the outcome measures and we prefer the greater variation afforded by this continuous variable.

[26] The unconditional effect on returning to teach in the same school is not significantly different from zero.

The variables we include the factor analysis are missing for some teachers. Traditional factor analysis fits the factor model to the correlation matrix constructed using only observations with complete data. In order to use all of the available data, we instead estimated the factor analysis using the pair-wise item correlation matrix. We apply a Promax rotation to the factor loadings. The resulting factors may be correlated with each other, but maximize the extent to which each measure is associated with a single factor. To choose the number of factors, we use an eigenvalue cut-off of one, a commonly used standard in this methodology.

The results of the factor analysis are reported in Table 7. The factor analysis results in two factors, which we call "cognitive skills" and "non-cognitive skills." The six variables with the largest positive loadings on the first factor are all reasonable proxies for cognitive skills: being a TFA corps member, attending a more selective college, SAT math score, SAT verbal score, cognitive ability as measured by the Raven IQ test, and math knowledge for teaching. The five variables with the largest positive loadings on the second factor are all reasonable proxies for other non-cognitive 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) have considerable *negative* loadings on the non-cognitive factor, while majoring in education has a considerable negative loading on the cognitive factor.

The measures that primarily load on a single factor are noisy estimates of that factor. In this case, the square of the loading coefficient reported in Table 7 is equal to the measure's reliability as an estimate of the underlying factor (the percent of the total variance in the measure due to the factor). Thus, the six measures that have loadings on the cognitive skills factor of around 0.6 have reliability as measures of cognitive skills of around 36%. Simply averaging

across these six measures would reduce the noise by 1/6, and increase the reliability to around 80%. A similar calculation for the 6 measures with loadings above 0.35 for the non-cognitive factor (including the negative of teaching fellow) increases the reliability from around 20% for any individual measure to over 60% for the average of the six measures.

We use the results of the factor analysis to predict each factor using all of the information available on each teacher. Most of these teachers only had a subset of the measures that were included in the factor model reported in Table 7, but the structure of the factor model allowed us to predict the underlying factors conditional on whatever measures were available. These predictions are linear combinations of all the measures from Table 7, and are the best linear unbiased predictor of the underlying factors. Therefore, they are in the same units as the underlying factor (which are normalized to have standard deviation equal to one). In total, we are able to measure these factors for a total of 403 teachers. We present results using the predictions from the factor model as measures of teachers' cognitive and non-cognitive skills because this is the more standard approach in the use of factor analysis. However, these predictions are highly correlated with the simple average of the six 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. Unlike for some of our non-traditional predictors, we do <u>not</u> standardize the factors to have a mean zero and standard deviation equal to one. Thus, the coefficients are indicative of a 1 point change in the underlying factor. It thus reflects our best estimate of the impact of a one standard deviation of cognitive or non-cognitive skills in the population of new teachers, not solely among survey respondents. Both factors are positively

and significantly associated with math achievement. Increasing either cognitive or non-cognitive skills by one point is associated with increases in student achievement of 0.033 standard deviations. Interestingly, only non-cognitive skills have a significant positive relationship with subjective evaluations, while cognitive skills have a significant positive association with retention within the DOE.

The effects of cognitive and non-cognitive 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 non-traditional 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 non-cognitive factors using only the traditional credentials (i.e., we act as if the non-traditional measures were unavailable for our survey respondents), repeat our regression analysis, and again predict impacts for respondents.[27] We then plot the distributions for these two sets of estimates in Figure 1. For additional comparisons, we also plot a simulated distribution of teacher effectiveness, which is simply a normal distribution with a standard deviation of 0.10. This is approximates the variation in value-added among new teachers estimated by Kane et al. (2006) for New York City teachers and serves as a simple benchmark against which to measures 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 non-traditional credentials (Figure 1). The standard deviation of predicted teacher effectiveness using only the traditional credentials to generate our factor estimates is 0.021, and adding the non-traditional credentials raises the standard deviation

[27] As we would expect, the coefficients in this additional regression are nearly identical (0.033 for cognitive skills and 0.034 for non-cognitive). However, the variation in the factors decreases due to the smaller number of variables used to make the factor estimates.

to 0.035.[28] 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 12 percent of the variance of the expected distribution of teacher effectiveness. This underscores the difficult, perhaps impossible, task of identifying systematically the most highly effective or ineffective teachers without any data on actual performance in the classroom.

7. 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 broadened set of information on new recruits. The evidence we present suggests that this is the case, and 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 paper by Boyd et al. (2008a) which makes the argument that recruiting teachers with a number of attractive credentials while avoiding teachers whose credentials are unattractive has 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 rely 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.

[28] The bimodal distribution of predicted effectiveness based on traditional characteristics 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 of 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 to 15) generates substantial variance in "predicted effectiveness," on the order of 0.06 to 0.08 standard deviations.

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.

References

Aaronson, D., Barrow, L. & Sander, W. (2007) "Teachers and Student Achievement in the Chicago Public High Schools," Journal of Labor Economics, 25(1): 95-135.

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E. & Zellman, G. (1976) Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools. Santa Monica, CA: Rand Corp.

Bandura, A. (1977) "Self-efficacy: Toward a Unifying Theory of Behavioral Change," Psychological Review 84(2): 191-215.

Barrick, M. R. and Mount, M. K. (1991) "The Big Five Personality Dimensions and Job Performance: A meta-analysis," Personnel Psychology, 44(1), 1-26.

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

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., and Wyckoff, J. (2008a) "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High-Poverty Schools", NBER Working Paper #14021

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2008b) "The Impact of Assessment and Accountability on Teacher Recruitment and Retention Are There Unintended Consequences?" Public Finance Review, Vol. 36(1): 88-111.

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L. "Teacher-Student Matching and the Assessment of Teacher Effectiveness" Journal of Human Resources 41(4):778-820 (2006)

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L. (2007) "How and Why Do Teacher Credentials Matter for Student Achievement?" NBER Working Paper 12828

Decker, P.T., Mayer, D.P. and Glazerman, S. (2004) "The Effects of Teach For America on Students: Findings from a National Evaluation," Mathematica Policy Research Report No. 8792-750.

Dembo, M.H. and Gibson, S. (1985) "Teachers' Sense of Efficacy: An Important Factor in School Improvement" The Elementary School Journal, 86(2): 173-184.

Ducharme, R. J. (1970) "Selected Pre-service Factors Related to Success of the Beginning Teacher," Doctoral Dissertation Louisiana State and Agricultural and Mechanical College.

Ferguson, R.F. and Ladd, H.F. (1996) "How and why money matters: An analysis of Alabama schools," in Ladd, H.F. Holding Schools Accountable: Performance-Based Reform in Education (p. 265-298). Washington, D.C: The Brookings Institution.

Getzels, J. W., and Jackson P. W. (1963) “The Teacher's Personality and Characteristics,” in N. L. Gate (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally.

Gibson, S. and Dembo, M.H. (1984) “Teacher Efficacy: A Construct Validation” Journal of Educational Psychology 76(4): 569-582

Goldhaber D. (2007) “Everyone’s Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?” Journal of Human Resources; 42(4): 765-794

Goldhaber, D. and Brewer, D. (1997) “Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity” Journal of Human Resources 32(3): 505-523.

Goodstein, L. D., and Lanyon, R. I. (1999). Applications of Personality Assessment to the Workplace: A Review. Journal of Business and Psychology, 13(3), 291-322.

Gordon, R., Kane, T., & Staiger, D. (2006) The Hamilton Project: Identifying Effective Teachers Using Performance on the Job. Washington, DC: The Brookings Institution.

Greenwald, R., Hedges, L.V. & Laine, R.D. (1996) “The Effect of School Resources on Student Achievement,” Review of Educational Research, 66(3): 361-396.

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

Haberman, M. (1995). Selecting "Star" Teachers for Children and Youth in Urban Poverty. Phi Delta Kappan, 76(10), 777-781.

Hanushek, E.A. (1986) “The Economics of Schooling: Production and Efficiency in Public Schools,” Journal of Economic Literature, 24(3): 1141-1177.

Hanushek, E.A. (1997) “Assessing the Effects of School Resources on Student Performance: An Update,” Educational Evaluation and Policy Analysis, 19(2): 141-164.

Harris, D.N. & Sass, T.R. (2006) “The Effects of Teacher Training on Teacher Value Added,” Working Papers wp_2006_03_01, Department of Economics, Florida State University.

Harris, D.N. and Sass, T.R. (2008) “What Makes for a Good Teacher and Who Can Tell?” Unpublished manuscript, Florida State University.

Hill, H. C., Rowan, B., and Ball, D. L. (2005) “Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement,” American Educational Research Journal, 42(2): 371-406.

Hill, H. (2006). Content Knowledge for Teaching Mathematics Measures (CKTM measures): Introduction to CKT-M scales: University of Michigan.

Hill, H. C., Rowan, B., and Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. American Educational Research Journal, 42(2), 371-406.

Hoy, W.K. and Woolfolk, A.E. (1993). Teachers' sense of efficacy and the organizational health of
schools. The Elementary School Journal 93, 356-372.

Jacob, B.A. (2007) "The Challenges of Staffing Urban Schools with Effective Teachers," The Future of Children 17(1): 129-153.

Jacob, B.A., and Lefgren, L.J. (2008) "Principals as Agents: Subjective Performance Measurement in Education" Journal of Labor Economics 26(1): 101-136.

John, O.P., Donahue, E.M., and Kentle, R. L. (1991). The "Big Five" Inventory—Versions 4a and 54. Berkeley: University of California, Berkeley, Institute of Personality and Social Research.

Kane, T. J., Rockoff, J. and Staiger, D. O. (2006). What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City. National Bureau of Economic Research Working Paper 12155.

Lachman, M.E. and Weaver, S.L. (1997) "The Midlife Development Inventory (MIDI) Personality Scales: Scale Construction and Scoring Technical Report – July 1997," Brandeis University Psychology Department, MS 062.

Maguire, J.W. (1966) "Factors in Undergraduate Teacher Education Related to Success in Teaching," Doctoral Dissertation, Florida State University.

Metzger, S. and Wu, M.J. (in press). Commercial Teacher Selection Instruments: The Validity of Selecting Teachers Through Beliefs, Attitudes, and Values. Review of Educational Research.

Murnane, R. J. (1975) The Impact of School Resources on the Learning of Inner City Children. Cambridge, MA: Ballinger.

Raudenbush, S.W., Rowan, B. and Cheong, Y.F. (1992) "Contextual Effects on the Self-perceived Efficacy of High School Teachers" Sociology of Education, 65(2): 150-167.

Raven, J. C., and Summers B. (1986). Manuel for Raven's Progressive Matrices and Vocabulary Scales - Research Supplement No 3. London: Lewis.

Raven, J. C., Court, J. H., and Raven, J. (1983). Manuel for Raven's Progressive Matrices and Vocabulary Scales (Section 3) - Standard Progressive Matrices (1983 ed.). London: Lewis.

Raven, J. C., Court, J. H., and Raven, J. (2000). Manual for Raven's Progressive Matrices and Vocabular Scales (Section 3) - Standard Progressive Matrices (2000 ed.). San Antonio: Harcourt.

Raven, J.C. (2000) “The Raven’s Progressive Matrices: Change and Stability over Culture and Time,” Cognitive Psychology 41(1): 1-48.

Rivkin, S., Hanushek, E. A. & Kain, J. (2005) “Teachers, Schools, and Academic Achievement,” Econometrica, 73(2): 417–458.

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

Rockoff, J.E. (2008) “Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City” NBER Working Paper 13868.

Shulman, L.S. (1986) “Those Who Understand: Knowledge Growth in Teaching,” Educational Researcher, 15(2), 4-14.

Shulman, L.S. (1987) “Knowledge and Teaching: Foundations of the New Reform,” Harvard Educational Review, 57(1), 1-22.

Washburne, C. and Heil, L. M. (1960) “What Characteristics of Teachers Affect Children's Growth?” School Review, 68, 420-428.

Wilson, S. M., Shulman, L. S., and Richert, A. (1987) “150 Different Ways of Knowing: Representations of Knowledge in Teaching,” in J. Calderhead (Ed.), Exploring Teachers' Thinking Sussex, England: Holt, Rinehardt, and Winston.

Woolfolk, A. E., and Hoy, W. K. (1990). Prospective teachers' sense of efficacy and beliefs about control. Journal of Educational Psychology, 82(1), 81-91.

Table 1: Comparison of Teachers by Survey Invitation and Response

	Respondents	Non-Respondents	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%
Masters 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 a teacher was 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.

Table 2: Summary Statistics on Survey Responses

Academic Background	Observations	Mean	S.D.
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
Barrons Rank of College (1 to 9 scale, 1 is best)	248	5.6	1.9
Passed the LAST Certification Exam on 1st 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

Table 3: Traditional Predictors of Teacher and Student Outcomes

	Math Achievement	Subjective Evaluation	Teacher Absences	Returned to NYC	Returned to School | NYC
Credentials					
Has a Graduate Degree	-0.014	0.133	0.019	-0.050	-0.016
	(0.024)	(0.138)	(0.412)	(0.033)	(0.035)
	[0.557]	[0.338]	[0.962]	[0.130]	[0.649]
Passed LAST Certification Exam on 1st Attempt *(1=yes)*	0.035	0.123	0.013	-0.053	0.001
	(0.039)	(0.196)	(0.688)	(0.040)	(0.059)
	[0.369]	[0.529]	[0.984]	[0.185]	[0.982]
Teaching Fellow	-0.046	-0.184	1.006	0.118	-0.016
(Relative to Traditionally Certified)	(0.027)*	(0.138)	(0.514)*	(0.031)**	(0.040)
	[0.085]	[0.185]	[0.050]	[0.000]	[0.695]
TFA Corps Member	0.044	-0.052	-0.501	0.128	0.090
(Relative to Traditionally Certified)	(0.030)	(0.140)	(0.422)	(0.035)**	(0.035)**
	[0.151]	[0.710]	[0.235]	[0.000]	[0.011]
Math or Science Major	0.040	-0.048	-1.212	-0.063	-0.007
(Relative to Those Other Than Math, Science, or Education)	(0.031)	(0.183)	(0.529)**	(0.049)	(0.049)
	[0.2]	[0.795]	[0.022]	[0.201]	[0.879]
Education Major	-0.009	-0.117	-0.485	-0.097	0.041
(Relative to Those Other Than Math, Science, or Education)	(0.033)	(0.144)	(0.489)	(0.042)**	(0.038)
	[0.789]	[0.417]	[0.321]	[0.022]	[0.279]
Self-Reported SAT Math Score *(s.d.=1)*	0.012	0.004	-0.119	0.008	0.005
	(0.015)	(0.075)	(0.207)	(0.020)	(0.015)
	[0.41]	[0.960]	[0.564]	[0.686]	[0.715]
Self-Reported SAT Verbal Score *(s.d.=1)*	-0.003	0.035	0.145	0.026	0.004
	(0.014)	(0.081)	(0.228)	(0.020)	(0.022)
	[0.829]	[0.666]	[0.524]	[0.188]	[0.853]
Barrons Rank of College *(s.d.=1)*	0.022	-0.212	0.059	0.027	-0.014
	(0.012)*	(0.087)**	(0.217)	(0.018)	(0.022)
	[0.076]	[0.015]	[0.786]	[0.118]	[0.015]
Control for Student/School Characteristics and Zip Code FE	√	√	√	√	√
Observations	247,903	3,030	4,858	4,877	4,516

Note: Each set of coefficients (separated by dotted lines) represent different regressions. See text for a full listing of the student and school charateristics 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%;

Table 4: Non-Traditional Predictors of Teacher and Student Outcomes

	Math Achievement	Subjective Evaluation	Teacher Absences	Returned to NYC	to School | NYC
Cognitive Ability *(Percentile, s.d.=1)*	0.016	0.066	-0.422	0.016	0.021
	(0.012)	(0.058)	(0.227)*	(0.016)	(0.019)
	[0.174]	[0.254]	[0.063]	[0.315]	[0.270]
Math Knowledge for Teaching *(Percent Correct, s.d.=1)*	0.028	0.014	-0.407	0.006	-0.011
	(0.012)**	(0.065)	(0.208)*	(0.014)	(0.017)
	[0.024]	[0.828]	[0.051]	[0.659]	[0.504]
Conscientiousness *(s.d.=1)*	0.011	0.188	0.185	-0.000	0.010
	(0.011)	(0.059)**	(0.169)	(0.013)	(0.020)
	[0.319]	[0.001]	[0.273]	[0.982]	[0.624]
Extraversion *(s.d.=1)*	0.007	0.216	0.086	0.000	0.022
	(0.011)	(0.062)**	(0.189)	(0.015)	(0.017)
	[0.519]	[0.001]	[0.650]	[0.986]	[0.201]
General Efficacy *(s.d.=1)*	0.017	0.019	-0.028	0.037	0.009
	(0.012)	(0.057)	(0.192)	(0.016)**	(0.016)
	[0.149]	[0.736]	[0.885]	[0.024]	[0.591]
Personal Efficacy *(s.d.=1)*	0.012	0.192	0.148	0.015	0.014
	(0.011)	(0.060)**	(0.204)	(0.013)	(0.015)
	[0.271]	[0.001]	[0.470]	[0.280]	[0.372]
Control for Student/School Characteristics and Zip Code FE	√	√	√	√	√
Observations	247,903	3,030	4,858	4,877	4,516

Note: Each set of coefficients (separated by dotted lines) represent different regressions. See text for a full listing of the student and school charateristics 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%;

Table 5: Predictors of Performance on the Haberman Pre-Screener

	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 on 1st Attempt *(1=yes)*	0.167	0.352
	(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	-0.094	-0.005
(Relative to Majors Other Than Math, Science, or Education)	(0.068)	(0.174)
Education Major	0.005	0.006
(Relative to Majors 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)
Barrons Rank of College *(s.d.=1)*	0.029	0.069
	(0.032)	(0.071)
Non-Traditonal Credentials		
Cognitive Ability *(Percentile, s.d.=1)*	0.017	0.255
	(0.025)	(0.060)**
Math Knowledge for Teaching *(Percent Correct, s.d.=1)*	0.049	0.198
	(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	√	√

Note: Each set of coefficients (separated by dotted lines) represent 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 shown in parentheses. * significant at 10%; ** significant at 5%;

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	0.033	0.243	0.928	0.009	-0.064
	(0.031)	(0.175)	(0.564)	(0.035)	(0.050)
	[0.297]	[0.167]	[0.100]	[0.793]	[0.206]
Haberman Total Score *(s.d.=1)*	0.021	0.141	0.135	0.027	-0.040
	(0.013)	(0.065)**	(0.230)	(0.018)	(0.020)**
	[0.110]	[0.029]	[0.556]	[0.125]	[0.043]
Controls for School Charateristics and School Zip Code	√	√	√	√	√
Observations	244,235	2,970	4,754	4,773	4,421

Note: Each set of coefficients (separated by dotted lines) represent different regressions. All regressions include indicators for grades taught (for teachers who can be linked to student data), school level (primary, middle, high school, or other) and highest grade, school zip code fixed effects, and school level observable characteristics (percentage of students by gender, ethnicity, free lunch receipt, school eligibility for Title I, and the pupil-teacher ratio.). Standard errors (in parentheses) are clustered by school. * significant at 10%; ** significant at 5%;

Table 7: Factor Analysis on Predictor Variables

Item	Factor 1: Cognitive Skills	Factor 2: Non-Cognitive 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
Barrons 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

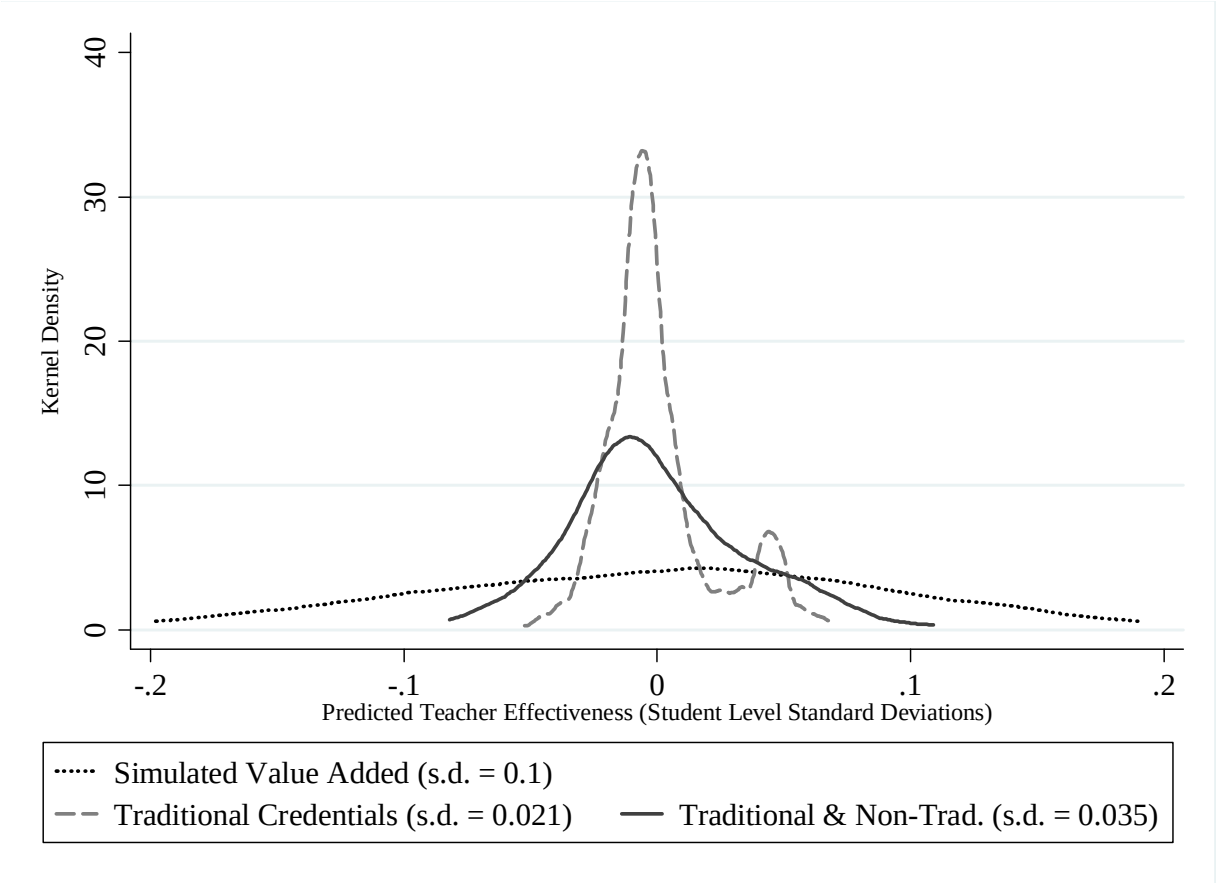
Notes: Factor loadings calculated using the pairwise item correlation matrix and applying a Promax rotation.

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

	Math Achievement	Subjective Evaluation	Teacher Absences	Returned to NYC	Returned to School | NYC
Factor 1: Cognitive Skills (s.d.=1)	0.033	0.025	-0.227	0.043	0.005
	(0.011)**	(0.065)	(0.195)	(0.016)**	(0.015)
Factor 2: Non-Cognitive Skills (s.d.=1)	0.033	0.272	-0.026	0.009	0.031
	(0.015)**	(0.068)**	(0.243)	(0.017)	(0.020)
F-Test: All Factors Equal Zero (p-value)	0.0023	0.00	0.51	0.03	0.30
Observations	247,903	3,030	4,858	4,877	4,516
Control for Student/School Characteristics and Zip Code FE	√	√	√	√	√

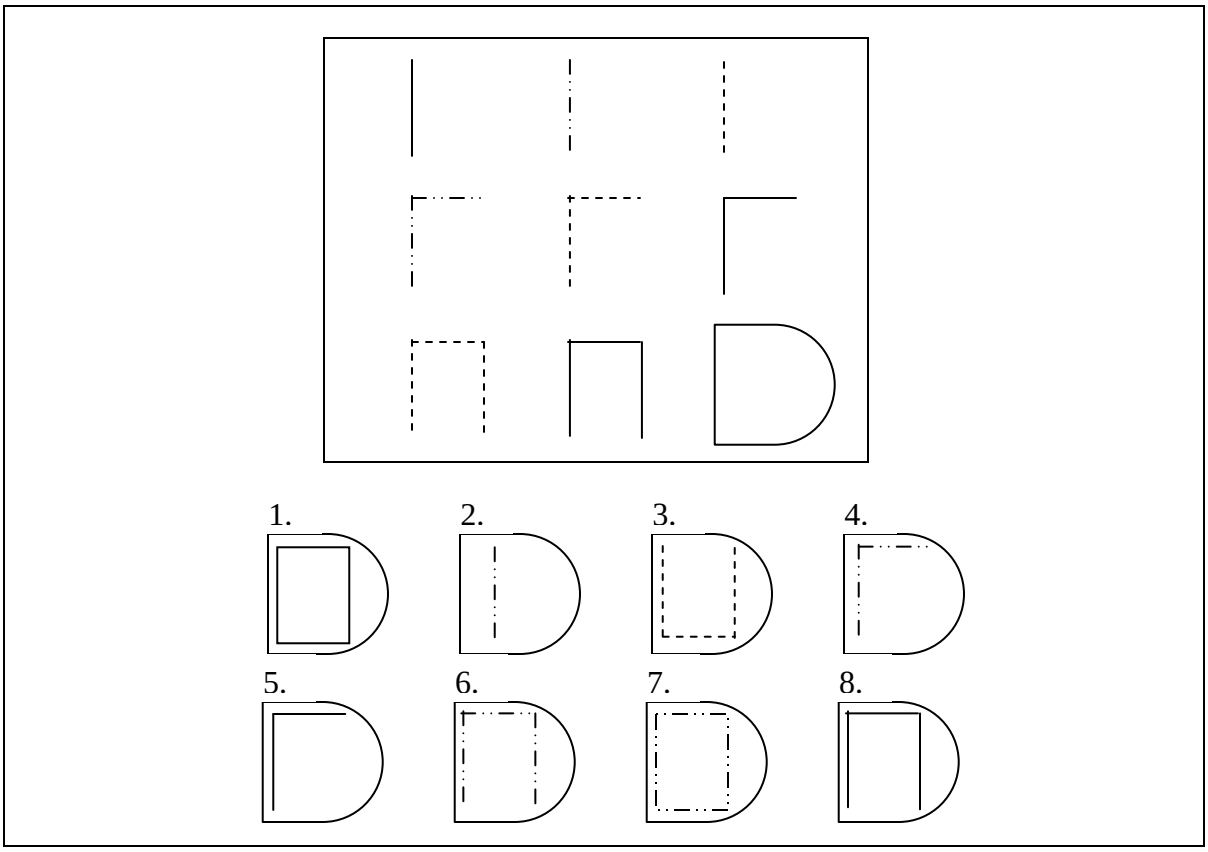
Notes: All regressions include grade level fixed effects, school zip code fixed effects, and student, class, and school level observable characteristics (see text for a complete list). Standard errors (in parentheses) are clustered at the school level. * significant at 10%; ** significant at 5%.

Figure 1: Recruitment Information and the Distribution of Predicted Value-Added



Note: Kernel density plots are shown of "Simulated Value Added" is the kernel density plot of a randomly drawn normally distributed variable with mean zero and standard deviation 0.10. The kernel density plots of predicted value-added from two regressions of student test scores on a set of teacher characteristics and other controls. 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).

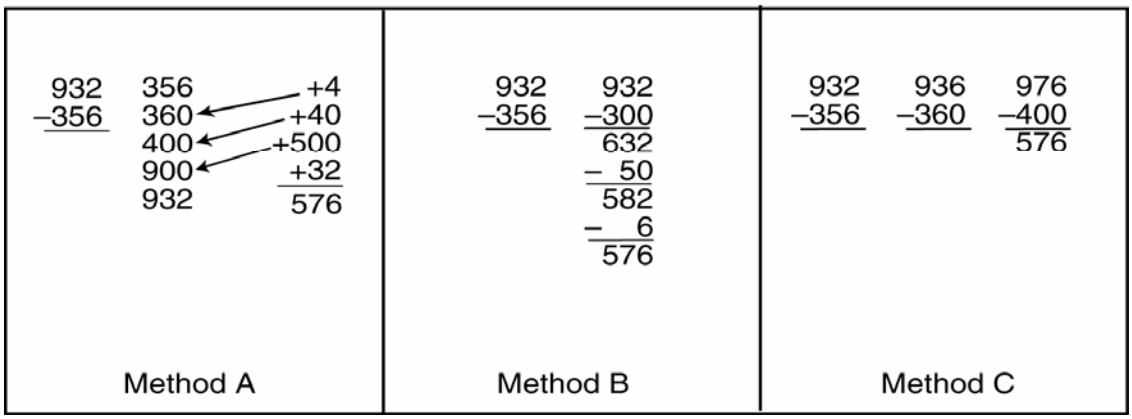
Appendix Figure 1:
Illustrative Item for Test of Cognitive Ability (Raven's Progressive Matrices)



Note: This item is reproduced from Figure 1 in Raven (2000). It is not from any currently used form of the Raven's Progressive Matrices test; it only illustrates the format of the test items. The design in a box at the top of the figure has a part missing, and test takers must select among the eight options below to complete the design. Although Raven (2000) does not give the correct response to this item, we surmise that it is option 6.

Appendix Figure 2:
Example of a Question on the Math Content Knowledge Test

Imagine that you are working with your class on subtracting large numbers. Among your students' papers, you notice that some have displayed their work in the following ways:



Which of these students is using a method that could be used to subtract any two whole numbers? (Select ONE answer.)

a) A only
b) B only
c) A and B
d) B and C
e) A, B, and C

Note: This item is taken from the Elementary Math section of the Math Content Knowledge Test. The correct response is (e).