CLOSUP
Center for Local, State, and Urban Policy
University of Michigan

CLOSUP Working Paper Series
Number 11

February 2009

Can You Recognize an Effective Teacher When You Recruit One?

Brian A. Jacob
University of Michigan

Thomas J. Kane
Harvard Graduate School of Education

Jonah E. Rockoff
Columbia Business School

Douglas O. Staiger
Dartmouth College

This paper is available online at http://closup.umich.edu

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the Center for Local, State, and Urban Policy or any sponsoring agency

Center for Local, State, and Urban Policy
Gerald R. Ford School of Public Policy
University of Michigan

Can You Recognize an Effective Teacher When You Recruit One?

Brian A. Jacob
University of Michigan

Thomas J. Kane
Harvard Graduate School of Education

Jonah E. Rockoff
Columbia Business School

Douglas O. Staiger
Dartmouth College

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: teaching specific content knowledge, cognitive ability, personality traits, feelings of self-efficacy, and scores on a commercially available teacher selection instrument. We find that a number of these predictors have statistically and economically significant relationships with student and teacher outcomes. 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.

* The authors would like to first thank John Fullerton, who helped us greatly in the design and implementation of this survey. We also thank a number of individuals who made this 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.

"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.[1]

[1] We have been unable to find the exact source of this commonly used anecdote, but it dates back at least to the 13th century Turkish stories of Nasreddin, who, in that version, lost his wedding ring in the darkness of his house but searched for it outside in the moonlight (Farzad (1989)).

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 and 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 significant relationships between student achievement and non-traditional predictors of teacher effectiveness, including performance on the Haberman selection instrument. We find positive effects of cognitive ability and self-efficacy which are about 0.02 standard deviations of student achievement and marginally significant. Stronger effects are found for respondents’ scores on a test of math knowledge for teaching, with an effect size of about 0.03 standard deviations and statistical significance at the 1 percent level. Scores on the Haberman PreScreener are also positively related to student achievement, with an effect size of 0.023 standard deviations which is significant at the 8 percent level. Interestingly, we do not find respondents’ levels of conscientiousness or extraversion (as measured on a standard personality inventory) are 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 test has the ability to separate very large differences in teacher effectiveness among our survey respondents. However, we document that using a number of metrics together may have meaningful power for screening effective teachers at the time of hire. 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 seek out information outside the realm of traditional teaching credentials.

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, tests of the predictive power of these factors, and tests for heterogeneity in these effects. Section 7 provides a graphical illustration of the combined power of information on respondents to predict student achievement, and Section 8 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.).[2]

[2] 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

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 (2006), 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."[3] New teachers in New York State generally take three certification tests. The Liberal Arts and Science Test (LAST) is required for certification in all subjects, while teachers must also take the Assessment of Teaching Skills (ATS-W) and a Content Specialty Test (CST) depending on their subject area and certification. The ATS-W is not required of alternatively certified teachers (e.g., TFA and Teaching Fellows) and is therefore not reported for a many of the teachers who took our survey. All of the tests are graded on the same scale and have the same passing score, and teachers in New York can take certification tests an unlimited number of times until they pass. Boyd et al. (2006, 2008a, 2008b) use whether a teacher that passed the LAST on the first attempt as a marker of effectiveness, and we ask this directly of teachers. A few teachers reported passing the LAST on their first attempt but also reported a failing score; we recode those teachers to having failed the exam on their first attempt and we use their reported (failing) score in our analysis.

us. For a few colleges for whom the Barron's rating was missing, we filled in the information using Barron's ratings from 1984.

[3] 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.

In addition to certification exam performance, we asked teachers to report their scores on college entrance examinations. This information is not collected by the DOE, nor do we believe it is likely to be used by administrators in making hiring decisions. 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 well cited study by Ladd and Ferguson (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 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

[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.

(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 Prior Professional Experience

Teaching experience has consistently been found related to teacher effectiveness (see Jacob (2007) for a review). Similarly, one might hypothesize that other professional experience, particularly in fields involving children, produces valuable human capital for teaching. The administrative data we collect provides a limited measure of prior teaching experience, as captured in the individual's salary step, but this does not incorporate information on other potentially relevant experience.[5] We therefore asked as a series of questions on prior experience in our survey. First, we asked about any occupations the respondent had prior to teaching. Second, we asked teachers to report both paid and volunteer experience in the following fields: teaching (in New York, outside of New York, and overseas), 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. For each of these categories, respondents also provided information on the months worked, number of hours per week worked, whether they were paid (or volunteered), and whether the work took place in a school.

The dimensionality of this information is too large for us to incorporate it all in any systematic way. We therefore use these responses to construct two broad measures of prior

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)).

[5] Moreover, the vast majority of new teachers hired by New York City and other high poverty districts will have no prior teaching experience whatsoever.

experience. The first is the number of hours of prior experience in a teaching role (i.e., as a teacher, a substitute teacher, or a paraprofessional) and hours of experience in other educational roles outside of teaching (i.e., as a tutor, employee in an after-school or community program, coach, camp counselor, mentor, or working in religious education). We calculate hours using the information on months worked and hours worked per week.[6]

2.3 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. 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

[6] Months worked was given categorically: 0-6, 7-12, 13-24, 25-36, or 37+. For each category, we assign numeric values of 12, 28, 52, 100, or 240 weeks. We cap total hours in teaching or educational non-teaching activities at 10,000. In addition, to these two variables on prior experience, we tried using an indicator for whether the teacher said that teaching in New York was their first occupation and the years of professional experience (regardless of occupation) for teachers who said that teaching in New York was not their first occupation. We did not find significant relationships between these variables and our outcomes of interest.

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.

2.4 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.

2.5 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 oftoverlooked 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.[7]

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.

2.6 Importance of Standardized Tests

One concern in the literature on teacher effectiveness is that teachers who produce large gains in standardized tests may do so at the detriment to other skills (see Koretz (2002) for a broader discussion and Carrell (2008) for recent empirical evidence). If this were the case, then one might hypothesize that teachers who support standardized tests or emphasize test performance skills might cause students to have higher test scores, but not perform better on other measures of teacher effectiveness (e.g., subjective evaluations). To this end, we asked teachers to state, on a 5-point scale, how much emphasis they place on various goals for their students. Based on these questions, we construct a measure of the *relative* emphasis teachers place on student performance in basic reading and math tests. Specifically, for each teacher, we take the average of his or her responses to two items (i.e., the emphasis placed on "basic math

[7] 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.

and reading skills" and "test taking skills") and subtract the average of the teacher's response to four other items (i.e., emphasis placed on "ability to work well with others", "love of learning", "personal growth", and "citizenship"). We rely on this relative measure in order to account for the fact that certain teachers may simply report placing more emphasis on everything. Second, we asked teachers specifically the degree to which they believe that New York State standardized tests accurately measure students' knowledge and skills. We generate an indicator for teachers that reported that the state tests measured students' knowledge and skills "very well" or "moderately well" (as opposed to "not well" or "not at all").

2.7 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' attributes.[8] 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

[8] 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.

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 little or no empirical data are available for independent analysis. A recent meta-analysis (Metzger and Wu (forthcoming)) analyzes 24 studies of Gallup's earlier protocol (the Teacher Perceiver Instrument) and finds little evidence of predictive power, but most of these studies involved small samples and only one used student achievement growth as an outcome measure. However, studies conducted by researchers at the Gallup Organization find a significant positive correlation between TIA scores and principals' evaluations of new teachers (Wallwey (2002)) and a significant positive correlation between recently hired teachers' TIA scores and student achievement growth (Kirk (2007a, 2007b)).

New York City recently began requiring all applicants for teaching positions to take the TIA, and, in other work, we will assess 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 questions correctly, but had a “low” score in one of the ten categories; and 4) a bottom group that consists of teachers who (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.[9]

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. Note that this is not based on any recommendation of Martin Haberman or the Haberman Foundation.

2.8 Other Teacher Characteristics

In addition to the items described above, we also asked about teachers’ childhood settings (i.e., rural, suburban, urban, or foreign), K-12 education (public or private), and attendance of New York City public schools. We do not test these measures as predictors of teacher effectiveness, as we did not have prior belief or encounter any research suggesting that that they would be. Nevertheless, it is worth noting that 54 percent of respondents grew up in a suburban setting, 30 percent in an urban setting, 11 percent in a rural setting, and 5 percent in a foreign country. Seventy-eight percent of respondents attended public schools and 30 percent attended New York City public schools for some part of their K-12 education.

[9] 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.

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 single elementary class, while others taught math but to one or more classrooms of students in middle school grades.[10]

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.[11] 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

[10] 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.

[11] 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.[12] In Section 4, we compare respondents and non-respondents on a variety of observable characteristics.

3.2 Administrative Data

In addition to the responses 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, uncertified, 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 significant and growing literature demonstrates a significant relationship between objective measures of teacher performance and subjective evaluations of teacher quality made

[12] 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).

during a 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. This 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)). Unfortunately, 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.[13]

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

[13] 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 the 4,275 new teachers with no prior teaching experience that started in the school year 2006-2007, 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.[14] 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). Most of missing evaluations (26 percent of the sample) are due to teachers working in Empowerment schools which did not participate in the centralized mentoring program. While all

[14] 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.

of the remaining teachers were supposed to receive mentoring, only 83 percent actually did, though this in line with earlier program years (see Rockoff (2008)) and is likely due to administrative errors and late hiring.[15]

For our analysis of student achievement, we use a slightly different sample. Specifically, we include all students and teachers in the value-added grades 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. (2007): 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.

[15] 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 21 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). BRIAN says: we should do a joint test of significance for all of these bivariate relationships that accounts for correlation among the predictors. i have some code to do this, so i can take care of it. 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.[16]

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,

[16] 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.

similar to 6.5 percent for non-respondents and 7.4 percent for other new teachers. An additional 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.[17] 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. 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. 2007), 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

[17] 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).

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.[18] 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) 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).

Over half of the survey respondents (54 percent) said that they had prior experience in a teaching role, and two thirds (66 percent) said that they had prior experience in an educational role other than teaching. Conditional on having some experience, the average total hours was higher for teaching (1,428 hours) than other roles (827 hours) and was roughly the equivalent of a year of full-time teaching (i.e., 180 days at 8 hours per day).

[18] As a basic 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 (scale of 1-5) on 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 (ICPSR Study No. 2760), 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. However, while this survey did use a similar format to our survey, the exact number and wording of the items were not identical and responses were given on a scale of 1 to 4 (see Lachman and Weaver (1997)), so the two sets of results are not directly comparable. 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

[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).

Due to the normalization described above, the average value for emphasis placed on test performance is close to zero. The mean value of 0.08 means that survey respondents were slightly more likely to report greater emphasis on test performance skills than “soft” skills such as “love of learning” and “ability to work well with others.” Roughly half of the respondents (49 percent) felt that state standardized tests were good measures of students’ knowledge and skills. Interestingly, the correlation between the measures of teaching emphasis and feelings towards the standardized tests is small and statistically insignificant (-.02).

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. Recall that 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 whether 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 control 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 the interpretation of our estimates. Three issues are worth noting. 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. 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

[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.

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, this bias our results towards zero. Brian says: it would be nice if we could think of a way to test the severity of this concern; perhaps some back-of-the-envelope bounding exercises? maybe tom or doug will have some thoughts here.

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., teacher certification score, 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.

5.1 The Power of Individual Predictors of Teacher Effectiveness

Note to coauthors: I've dropped the experience variables from the discussion (and they're hidden in the regression table). For a while I've felt like we were grasping for straws on these things and didn’t really have any clear rationale for the variables we constructed. Also, the three permutations we’ve tried have basically gone nowhere. If you want to add them back then we need to add these as results in the discussion here, but I would prefer to relegate the professional experience stuff to a footnote/paragraph that we tried various things and nothing came up.

Table 3 shows results for the power of teacher demographics and 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. Using demographics variables, we find, among our survey respondents, female respondents and younger respondents were, on average, more effective than their male and older colleagues at raising achievement, while teacher effectiveness is not significantly related to ethnicity.[21] These results are similar if we use all teachers to identify the coefficients, as opposed to just survey respondents. While these findings are interesting, legal constraints do not permit schools to use demographics in making hiring decisions.

The second part of Table 3 presents results on a set of traditional hiring credentials for new teachers. 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 LAST exam 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.

[21] For simplicity of exposition, we show the coefficient on whether a teacher is white vs. non-white, but we find no significant variation across ethnicities when including indicators for whether a teacher is black, Hispanic, Asian, or Native American.

When comparing alternatively certified teachers to traditionally certified among the survey respondents, we find that teaching fellows are significantly less effective (0.05 standard deviations) and Teach for America corps members are more effective, although this latter coefficient is only marginally significant (p-value = 0.15).[22] While the result on TFA is consistent with other findings (Decker et al. (2004), Boyd et al. (2006), Kane et al. (2008)), the negative finding for teaching fellows contrasts with earlier work (Boyd et al. (2006), Kane et al. (2008)). However, in contrast to the coefficients on demographics for our survey respondents, the negative finding on Teaching Fellows disappears when we use identifying variation in the certification pathway of all teachers. Thus, it appears to be the case that 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, Kane et al. (2008) present evidence in support of the latter hypothesis.[23]

Students' test scores growth was greater on average with respondents who majored in math or science and lower with respondents who majored in education, but neither coefficient is statistically significant nor can we 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 = 0.11). The positive, albeit small, relationship between college selectivity and teacher effectiveness has been found in other studies (Clotfelter et al. (2007), Boyd et al. (2008) others?). The lack of statistical significance for SAT scores contrasts with findings from other research, but it is worth pointing

[22] 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.

[23] Non-random selection of survey respondents does not drive this result. The coefficient does not change when we use identifying variation on all teachers who were asked to take the survey, as opposed to only respondents.

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 two variables among the demographics and traditional credentials that are related to subjective evaluations are ethnicity (higher evaluations for white respondents) and college selectivity (lower evaluations for respondents that attended more selective colleges). Upon closer inspection, we find that the higher average evaluations among white respondents are driven by three factors: a significant difference between white and non-white respondents in the relationship between evaluations and the probability of survey response, a greater difference in evaluations among teaching fellows between white and non-white respondents, and particularly low evaluations for Asian respondents. Among non-white respondents, respondents had lower average evaluations (-0.13) than non-respondents (-0.03), while among white respondents the pattern was reversed (-.06 for non-respondents and 0.13 for respondents). Among teaching fellows, the difference between white and non-white respondents' evaluations was greater (0.36) than for respondents from other routes (0.2). In addition, the average evaluation for Asian respondents was -0.24. Thus, the overall difference in evaluations between white and non-white respondents' (0.27) is triple the difference found if we include non-respondents and exclude teaching fellows and Asian respondents (0.09).[24]

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 1.2 days more on average than other respondents.

[24] Recall that mentor ratings have been normalized by mentor, so none of these differences are caused by variation across mentors in leniency/harshness of the evaluations.

Female respondents also had higher rates of absences (0.9 days). No other demographic characteristics or teaching credentials were significant predictors of absences.

With regard to retention, we find that older respondents, respondents with graduate degrees, and education majors are less likely to return to teaching in the DOE the following year, and that teaching fellows and TFA corps members were more likely to return. These results support the notion that teachers with more outside job opportunities are more likely to leave teaching in New York. 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. Female respondents, conditional on returning to the DOE, are more likely to transfer to another school.

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 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 is positive are marginally significant (p-value = 0.15) with a coefficient of 0.017, suggesting that cognitive ability does bear some relation to teacher effectiveness. Math knowledge for teaching is more strongly related to math achievement, with a coefficient of 0.032 which is statistically significant at the 1 percent level. This gives substantial 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 (0.014) and extraversion (0.006) are positive, but their p-values (0.23 and 0.57, respectively) are not significant. For general and personal efficacy, we also find positive coefficients (0.019 and 0.011, respectively) with marginal significance on general efficacy (p-value = 0.12). Overall, these results give mild support to the notion that teachers' personality and attitudes are related to teacher effectiveness.[25]

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 0.17 to 0.21. In contrast, the evaluations bear little relation to the three 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 the overall relationship between subjective evaluations and students' math achievement is positive and both statistically and economically significant, so that at least a portion of the variation in evaluations is based on observable differences in teachers' abilities to raise student achievement. 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

[25] 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 Why students perform worse with teachers who believe the state tests are good measures of students' knowledge is unclear, 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.

with a 0.05 standard deviation increase in math test scores.[26] However, a portion of the variance in 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 the characteristics that correlate with evaluations may be unrelated to productivity.

None of these non-traditional predictors are significantly related to teacher absences. With regard to retention, respondents with high cognitive ability scores were more likely to return to the DOE and more likely to return to the same school the following year. Each of these effects are marginally significant (p-values = 0.17 and 0.16), but if we look at the unconditional probability of returning to the same school we find a coefficient of 0.04 with a p-value of 0.08. Teachers who expressed high general efficacy were also more likely to return to the DOE in the following school year. 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 influenced by feelings about self-efficacy. However, it is worth noting that the questions regarding personal efficacy, as opposed to general, 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.

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

[26] The use of subjective evaluations as a means for identifying effective teachers after the recruitment stage is the subject of other research by one of the authors.

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. Note to coauthors: the excel file for table 5 has columns hidden that (a) drop the control variables and (b) include all predictors at once.

Performance on the Haberman PreScreener is significantly related to a number of these variables (Table 5). White, and older respondents perform better, particularly with regard to making the top group. [27] Performance is also higher for respondents who passed the LAST on

[27] 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

their first attempt or 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. 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. 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 significant 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 deviation increase in math achievement. Increasing scores are also associated with higher subjective evaluations and a greater propensity to return to teaching the following year. 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 provide a straightforward characterization of the predictive power of various teacher characteristics. 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 more consistent predictive power for teacher and student

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.

outcomes. We conduct a factor analysis to see whether it is possible to construct a smaller number of variables to use as predictors in a simplified analysis. In the factor analysis, we include all of the variables whose coefficients are shown in Tables 3, 4, and 5, with few exceptions. We do not include teacher demographics or certification pathway, since these describe groups of individuals rather than particular traits or measures of ability. We also do not use the indicator for being in the top group of candidates according to the Haberman methodology and just include the total score. While the former is recommended by the Haberman Foundation, we have seen that the total score has a stronger relationship with the outcome measures and we prefer the greater variation afforded by this continuous variable.

The variables we include the factor analysis are missing for some teachers. We therefore conduct the factor analysis using the pair-wise item correlation matrix, and apply a Promax rotation to the factor loadings. Using an eigenvalue cut-off of one, the factor analysis results in two factors, which we call "cognitive skills" and "non-cognitive skills," and the 16 predictors load onto these factors in fairly predictable ways, as shown in Table 7. The variables which load onto cognitive skills are: majoring in math or science, passing the LAST exam on the first attempt, selectivity of college, reported SAT math and verbal scores, cognitive ability, and math knowledge for teaching. The variables that load onto non-cognitive skills are: majoring in education, having a graduate degree, prior teaching experience, extraversion, conscientiousness, and personal efficacy. The remaining variables— prior experience with kids, general efficacy, and Haberman total score—are roughly equally loaded onto both factors. We use the results of the factor analysis to predict each factor using all of the information available on each teacher. (XX Doug—we should put down more of the methodology/matrix algebra here XX.) In total, we are able to measure these factors for a total of 403 teachers (402 for “ff”, 396 for “fff”). For ease of interpretation, we standardize each factor to have a standard deviation equal to one.

We estimate regressions using the same specification as before but now focusing on just these two factors as predictors. The two factors are nearly orthogonal (r = -0.07) (0.02 ff 0.08 fff), and we present results from specifications that include both factors together. Both factors are positively and significantly associated with math achievement (Table 8). Increasing cognitive and non-cognitive skills by one standard deviation is associated with increases in student achievement of 0.027 and 0.021 standard deviations, respectively. These are modest but still economically important effects. To see this, suppose we take the estimates from Table 8 and assign each respondent the predicted impact on student achievement associated with these two variables, so that a candidate one standard deviation above the mean on both traits will have a predicted impact of 0.048. The difference between the 75th and 25th percentile respondent on this measure is 0.043 standard deviations, and the difference between the 90th and 10th percentile respondent on this measure is 0.078 standard deviations.

Turning to the teacher level outcomes, we find that non-cognitive skills are significantly related to subjective evaluations, while cognitive skills have a significant association with retention within the DOE. Consistent with earlier results, neither of these factors is significantly related to teacher absences or retention within the same school conditional on returning to teach in the DOE.

6.1 Heterogeneous Effects

A number of researchers have examined the hypotheses that individuals with particular characteristics may be better at teaching students with particular characteristics. One particularly well-identified study is Dee (2004), who finds that black and white students scored higher on tests of math and reading achievement when randomly assigned to a teacher of the same race. Others have examined similar issues in non-experimental settings (e.g., Clotfelter et al. (2006), Kane et al. (2006), Lockwood and McCaffrey (2008)).

We examine whether the predictive power of the cognitive and non-cognitive skills factors have different predictive power for different groups of students and schools. Specifically, we examine heterogeneity between elementary and middle school grade levels, high and low poverty schools, and students with relatively low and high prior achievement. We divide schools by poverty by taking schools with above and below 80 percent of students on free lunch. The median percent on free lunch for all schools is about 78 percent, though the schools that hired survey respondents have somewhat higher poverty rates (median 82 percent). To divide students by prior achievement, we measure whether a student's prior year achievement score in math was above zero; recall that achievement is normalized by grade and year to have a mean of zero. For simplicity, we only present results here for student achievement, the outcome of greatest interest.

The results of this analysis are shown in Table 9. Note to co-authors: I've hidden the rows with the FF and FFF factors, so we just need to swap them for the F predictors if we decide to change the factor analysis. The results did not change. We find little evidence of heterogeneity in the relationships between student achievement and these two factors for teachers working in elementary and middle school grades (Column 1) or for teachers working in high poverty and medium/low poverty schools (Column 2). The point estimates suggest that non-cognitive skills may be more important for elementary students and high poverty schools, but these interaction terms are not statistically significant. In contrast, the interaction between cognitive skills and an indicator for students whose prior scores were above average is highly significant and suggests that most of the overall effect of cognitive skills on achievement is driven by these students. Assignment to a teacher with high cognitive skills has a positive but small and statistically insignificant impact on students with prior scores below the mean.

7. Measuring the Potential Power of Information at Recruitment

In order to gauge the overall usefulness of the results presented here, we present a graphical illustration of the power of using an expanded set of information at recruitment. We take the estimates from regressions of the form used earlier, each with a different set of respondent characteristics included as predictors of student achievement. Then, using the coefficients on teacher characteristics, we calculate a predicted value-added for each respondent, and plot the distribution of these estimates to give a sense of the magnitude of the predictive power of the data. If the predicted distribution is wide, it means that we might expect large differences between teaching candidates with favorable characteristics and those with unfavorable characteristics. For purposes of comparison, we plot distributions using only traditional teaching credential, using only the non-traditional predictors from our survey (including the Haberman PreScreener), and using the two factors for cognitive and non-cognitive skills. In addition, we plot a simulated distribution of teacher value-added, using random draws from 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. (2007) for New York City teachers and serves as a simple benchmark against which to measures the variation in predicted teacher effectiveness using the administrative and survey information.

Plotting the resulting estimates, we see a clear increase in the variation of predicted teacher effectiveness as we use more information (Figure 1). The standard deviation of predicted teacher effectiveness using only the traditional credentials is 0.055, and the standard deviation of predicted teacher effectiveness using only the non-traditional predictors from our survey is

0.039. Using both sets of variables raises the standard deviation to 0.067. This exercise suggests that districts may be able to gain significant traction in selecting more effective teachers by using broader sets of information during recruitment. However, it is important to stress the caveat that any 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. Thus, the validity of the predictions from our analysis can only be truly tested using data on a different sample of new teachers and their students.

8. Conclusion

We use a survey of new teachers in New York City to investigate whether economically significant variation in teacher effectiveness can be predicted 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 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. (2008) 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 any teachers' curriculum vitae.

While our findings provide substantial 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

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.

Dee, T.S. (2004) “Teachers, Race, and Student Achievement in a Randomized Experiment,” Review of Economics and Statistics 86(1): 195-210.

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.

Ehrenberg, R.G., and Brewer, D.J. (1994) “Do School and Teacher Characteristics Matter? Evidence from High School and Beyond,” Economics of Education Review 13(1): 1-17.

Farzad, H. (1989) Classic tales of Mulla Nasreddin. Costa Mesa, CA: Mazdâ Publishers. (translated by Diane L. Wilcox).

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.

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.

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, 200X

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.

Kirk, P.S. (2007a) "From Doubt to Belief in the TeacherInsight," Gallup Education Division Research Review.

Kirk, P.S. (2007b) "TeacherInsight Validation Using Achievement Data," Unpublished Manuscript.

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.

J.R. Lockwood Daniel F. McCaffrey "Exploring Student-Teacher Interactions in Longitudinal Achievement Data" RAND Corporation April 7, 2008

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.

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 B., S. (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.

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.

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.