Predicting Student Outcomes from Information Knowable at the Time of Hire: A **Systematic Review**

Jeffrey C. Valentine
Christopher R. Rakes
Dericka Canada

University of Louisville

DRAFT August 10, 2010
Please do not distribute without the permission of the first author.

Predicting Student Outcomes from Information Knowable at the Time of Hire: A **Systematic Review**

Executive Summary

This paper reports on a systematic review of studies that examine the relation between characteristics of teachers knowable at the time of hire relevant employment outcomes (most notably, student academic achievement). In this report, we discuss the difficulties inherent in assessing both important teacher candidate characteristics and student achievement, and in linking teacher candidate characteristics to student achievement outcomes. The context of this investigation is based on our experience working with Jefferson County Public Schools (JCPS).

For all of the political discussions regarding linking teacher pay to student outcomes, relatively little attention has been paid to the difficulties of doing so in a reasonable and believable way. Efforts to link student outcomes to teacher background characteristics face similar challenges. Virtually all studies that attempt to link teacher characteristics with student outcomes do so with easily measured variables. For example, student achievement is usually measured using standardized achievement tests, which, though objective, incompletely sample the construct of achievement and when incentivized are prone to score inflation. Furthermore, such measures of student achievement do not necessarily reflect the quality of student learning (e.g., Marzano, Pickering, & Pollock, 2001).

Student background characteristics, such as socioeconomic status, are also not often measured well (e.g., studies often use free lunch status as a proxy). Similarly, the study of teacher characteristics is usually limited to easily observed characteristics such as credentials, usually does not delve into more difficult to measure psychological constructs, and rarely involves structured observations of actual teaching or evaluations of teaching portfolios. The over-reliance on easily measured variables occurs in part because measuring important aspects of teacher behavior and student achievement is quite costly and often not well understood.

An additional difficulty involved in studying the relation between teacher characteristics and student outcomes is that students are almost never formally randomly assigned to teachers, and likewise teachers are seldom randomly assigned to schools. Especially given that inexperienced teachers are more likely than experienced teachers to be assigned to distressed schools, the latter concern means that it is important for researchers to identify, reliably measure, and take into account school contextual variables (e.g., the quality of the principal, SES characteristics of the children served) that might also influence student achievement. Similarly, the fact that students are usually not randomly assigned to teachers means that the method for gathering the most compelling evidence (i.e., the randomized experiment) for assessing the extent to which teacher characteristics (such as being a "highly qualified" teacher) are associated with student outcomes is not a particularly feasible approach and indeed may be

considered unethical in some cases. For example, it would be difficult to justify a study in which teachers were first judged to be either "high quality" or "low quality", and then students were randomly assigned to be with either a "low quality" or a "high quality" teacher; similar concerns would exist with assigning students to experienced vs. inexperienced teachers. The unlikelihood of a random assignment process suggests that in some or even many cases, student assignments may be made on a basis that biases estimates of teaching effectiveness.

Further, teachers are but one of several sources of influence on student achievement. Other largely independent factors such as student (e.g., motivation, academic ability), family (e.g., socioeconomic status, parental involvement), and school (e.g., climate, quality of the principal) variables also have the potential to play important roles in student achievement. As such, the proportion of variability in student scores that is attributable to teachers is likely to be small, and the proportion of variability in student scores attributable to teacher characteristics knowable at the time of hire even smaller, thus making such effects more difficult to detect and studies of those effects more sensitive to errors and biases.

A final concern, particularly related to our goal of identifying characteristics of teachers knowable at the time of hire that predict student achievement, is that it is very difficult (and often impossible) to obtain data on teacher effectiveness from candidates who were not hired by the district. Some potential candidates might be employed by

other districts – and their effectiveness is theoretically knowable even if logistically the information is difficult to obtain – but others may never be hired to work as teachers. If the hiring process generally works as intended (so that better teachers have a higher probability of being hired) then this will introduce an artifact (range restriction) that will serve to attenuate observed relations. In other words, characteristics of teachers that actually do predict student achievement will be more difficult to detect.

Taken together, these concerns suggest that, regardless of the specific aspects of teachers that are being studied, the most appropriate role for analyses linking teachers to student achievement data is in identifying teachers who might benefit from additional professional development. In essentially all cases the data will simply be too "muddy" to be useful as an important consideration in, for example, termination or tenure decisions.

Results of the Systematic Review

Despite these difficulties, several researchers have attempted to link teacher characteristics and outcomes such as student achievement. Our systematic literature search uncovered 14 studies that met our inclusion criteria. The studies assessed a variety of teacher background characteristics (e.g., credentials, knowledge, and academic preparation – such as participation in a program like Teach for America and the selectivity of the undergraduate institution attended). Most studies assessed the relation between teacher characteristics and student academic achievement, usually via

end of the year state-mandated tests. Though clearly much more research needs to be done, the studies in our review suggest that measures of teacher knowledge (such as scores on the SAT) might be related to student achievement. The data also suggest that candidates from the Teach for America program seem to achieve results that are at least as good as those to whom they were compared.

Future Directions for Districts

Given national trends, it seems sensible for districts to build information systems that will allow them to link and track a wide variety of types of information about teacher candidates, teachers, and students. Especially given some of the more aggressive recommendations for using student achievement data (such as in tenure and termination decisions), care will need to be taken to maximize the probability that such systems yield interpretable data. Existing systems in North Carolina, Florida, New York City, among others, could serve as potential models. That said, we reemphasize our skepticism that such information systems will yield data clean enough to support decisions like tenure and termination in specific cases. More defensible applications of analyses linking student achievement to teacher behaviors and characteristics include identifying (a) candidates for increased professional development and (b) characteristics of successful teachers that might inform hiring decisions.

The foregoing suggests that districts may wish to collect much more information on teacher candidates at the application stage, an activity that would be facilitated by

more research on the characteristics of teachers that seem to be associated with greater levels of student achievement. In addition, we strongly suspect that the analysis of student achievement will need to move beyond the end-of-the-year, state mandated tests employed in most studies. In part this concern is reflective of the need to conceptualize achievement more broadly. In addition, any decision based on mono-measurement models place too much emphasis on the single measurement. Multiple measures of different types are critical components for measurement reliability. In addition, most agree that individual teachers should be assessed in terms of student growth, but given summer learning loss this suggests that districts would be better served by testing in the Fall and in the Spring (so that growth can be measured more effectively). Given the widely held perception that there is too much testing (Barton, 1999), we suggest that a re-thinking of the way testing is currently implemented in most districts is needed (specifically, a move away from high-stakes testing and towards more diagnostic testing). That said, it does point out the need for informative data collection at national, state, and local levels. This study highlights some of the tradeoffs involved in trying to balance a realistic testing regimen with the need of policymakers, administrators, and parents to be able to identify the most effective teachers.

PREDICTING STUDENT OUTCOMES FROM INFORMATION KNOWABLE AT THE TIME OF HIRE: A SYSTEMATIC REVIEW

Like virtually all organizations that hire people to carry out their work, school districts have a strong interest in hiring candidates who are most likely to be successful. The "person on the street" definition of a successful teacher is one who teaches his or her students well. However, success is a multidimensional construct that also encompasses variables such as duration in the teaching profession (especially in the district in which hired) and collegiality, in addition to broader measures of academic achievement such as self-motivation and a instilling a love of learning. The focus of this paper is on the characteristics of teachers knowable at the time of hire and how well these predict student achievement, broadly defined. In it, we discuss the difficulties inherent in assessing both important teacher candidate characteristics and student achievement, and also in linking these to student achievement outcomes, and do so within the context of our experience working with Jefferson County Public Schools (JCPS). We conclude with a formal systematic review of studies examining this question, and suggest directions for future research.

A widely accepted proposition is that teachers are a major influence on the academic achievement of their students. In fact, of all the school context variables that

have been studied, teacher quality is arguably the one with the largest effects. This does not mean, however, that teacher quality itself has a large effect on student achievement. Rockoff (2004), for example, estimated that for elementary school students having a teacher who scores one standard deviation above the mean in teacher quality confers an advantage of 0.1 standard deviations on standardized mathematics and reading tests; similar findings were observed by Aaronson, Barrow, and Sander (2007) for high school students. While there may appear to be a discrepancy between the assertion that teachers are a major influence on achievement and the observed effect sizes, the elementary school effects are comparable to effects observed in studies of class size reduction (Konstantapoulous, 2008), and the high school effects are equivalent to perhaps half of a year's worth of expected academic growth in high school (Hill, Bloom, Rebeck Black, & Lipsey, 2008).

As noted, our specific interest with this research is to identify characteristics of teachers knowable at the time of hire that predict student learning. One such characteristic, teacher verbal ability, has typically been seen as one of the best predictors of student achievement. This notion was supported by data from the Coleman report (Coleman et al., 1966), though a recent systematic review and meta-analysis suggests that the relationship between teacher verbal ability and student achievement is not nearly as strong as once thought, and it is clear that more studies need to be conducted (Aloe & Becker, 2009).

A prototypical example of research that attempts to link teacher characteristics with student achievement was carried out by Rockoff, Jacob, Kane, and Steiger (2008). These researchers went well beyond typical studies by administering to new mathematics teachers an in-depth survey. This survey included measures of general cognitive ability, mathematics content knowledge, personality, and self-efficacy beliefs. The researchers also collected several relatively more traditional background measures such as self-reported scores on the SAT (mathematics and verbal), whether the teacher was an education major, and whether the teacher had a graduate degree. Finally, these researchers also used the Haberman Star Teacher Evaluation Prescreener (a commercial product that is intended to provide guidance to administrators regarding whether a particular candidate appears to be suited to teaching in urban environments; Haberman, 1993). As is typical in this line of research, Rockoff et al. also included a number of statistical control variables that attempt to reduce potential bias in their analyses; these included school level student ethnicity, gender, student-teacher ratio, and eligibility for free lunch, among others. Analyses of student achievement data also included controls for prior mathematics and reading test scores. They found that when individually considered, all teacher predictors had at best very small and generally not statistically significant associations with student outcomes. For example, having a new teacher who was a mathematics or science major was associated with better mathematics achievement, but this effect was small ($\beta = 0.04$) and not statistically

significant (p = .20). Similar results were observed for cognitive ability (β = 0.02, p = .17), the top group of Haberman scorers (β = 0.03, p = .30), and conscientiousness (β= 0.01, p = .52). Collectively, however, cognitive skills (measured by self-reported SAT scores, the cognitive ability test, and major, among others) and non-cognitive skills (such as self-efficacy, conscientiousness, and extraversion) were both associated with better student achievement, such that having a teacher who scored one standard deviation above the mean on either factor conferred an achievement advantage of about .03 units; both of these effects were statistically significant. Again, these effects are not large, but collectively they appear to be more important than any of the individual factors. If this finding is not simply an artifact (e.g., of greater reliability of the composite measure), it suggests that it may be a mistake to look for one single teacher background factor that contributes greatly to student achievement.

Metzger and Wu (2008) also authored a distinctive study about the relationship of teacher characteristics to student achievement. These researchers conducted a systematic review and meta-analysis of 24 studies that examined the Teacher Perceiver Inventory (TPI), a structured, face-to-face interview that is made up of 60 open-ended questions. Most studies examined the relation between scores on the TPI and ratings of teachers (from both administrators and students). These studies suggested fairly strong associations between scores on the TPI and ratings (r's around .25). Unfortunately, only one study examined the relation between scores on the TPI and student achievement

(this study found a positive association between those variables).

While the research studies discussed so far are notable, it must be made clear that examining the relation between teacher characteristics and student outcomes is not easy (this in part accounts for the paucity of research on the relation between the TPI and achievement). Among a host of other concerns, we believe five are most problematic. First, virtually all studies that link teacher characteristics and student outcomes do so with easily measured variables. This means that, for example, student achievement is usually measured using standardized achievement tests, which, though objective, incompletely sample the construct of achievement and when incentivized are prone to score inflation. Furthermore, measures of student achievement do not necessarily reflect the quality of student learning (Coleman, 1966; Marzano, Pickering, & Pollock, 2001). Student background characteristics, such as socioeconomic status, are also not often measured well (e.g., studies often use free lunch status as a proxy). Similarly, the study of teacher characteristics is usually limited to easily observed characteristics such as credentials, usually does not delve into more difficult to measure psychological constructs, and rarely involves structured observations of actual teaching or evaluations of teaching portfolios. The over-reliance on easily measured variables – which Rockoff et al. (2008) liken to the person looking for keys under a street light because that's where the light is – occurs in part because measuring important aspects of teacher behavior and student achievement is quite costly.

A second difficulty involved in linking teacher characteristics to student outcomes is that there is little agreement about which characteristics of teachers are important to measure, and even less agreement on the characteristics of a "quality teacher" (Berliner, 2005; Rice, 2003). Most often, measures of student achievement are used as proxies for teacher quality (Moyer-Packenham, Bolyard, Kitsantas, & Oh, 2008). In other words, good students are presumed to be the products of good teachers. For example, Phillips (2010) examined the effects of full state certification, teacher education levels, and content knowledge on student achievement. Others have examined outcomes such as teacher behaviors and knowledge commonly associated with student achievement (e.g., Scribner & Akiba, 2010). In rare cases, variables other than achievement such as student affect (e.g., Teven, 2007) and student learning (e.g., Chesebro, 2003) are used as proxies for teacher quality. Regardless of which proxy is used, the lack of a clear definition of high quality teacher characteristics makes measuring the teacher quality construct murky at best.

A third difficulty involved in studying the relation between teacher characteristics and student outcomes is that students are almost never formally randomly assigned to teachers, and likewise teachers are seldom randomly assigned to schools. Especially given that inexperienced teachers are more likely than experienced teachers to be assigned to distressed schools, the latter concern means that researchers must identify, reliably measure, and take into account school contextual variables (e.g.,

the quality of the principal, SES characteristics of the children served) that might also influence student achievement. Similarly, the fact that students are usually not randomly assigned to teachers means that the method for gathering the most compelling evidence (i.e., the randomized experiment) for assessing the extent to which teacher characteristics (such as being a "highly qualified" teacher) are associated with student outcomes is not a particularly feasible approach and indeed may be considered unethical in some cases. For example, a study in which teachers were first judged to be either "high quality" or "low quality", and then students were randomly assigned to be with either a "low quality" or a "high quality" teacher would be difficult to justify; similar concerns would exist with assigning students to experienced vs. inexperienced teachers. The lack of a random assignment process can lead to biasd estimates of teaching effectiveness. For example, Rothstein (2009, 2010) found that certain value-added models supported the conclusion that *fifth* grade teachers had an causal influence on *fourth* grade students' test scores. Thus high stakes evaluations based on measures not accounting for selection bias (Shadish, Cook, & Campbell, 2002) will "reward or punish teachers who do not deserve it and fail to reward or punish teachers who do" (Rothstein, 2010, p. 211).

How might selection bias reveal itself in classrooms? One example (Rothstein, 2010) suggested that it might occur when teachers attempt to sway principals and counselors to assign them honors classes, in which students might be expected to

perform higher than their peers in less advanced classes. It is also easy to imagine this selection bias coming from parents. For example, parents might push for their children to be taught by the more experienced, better credentialed, more highly thought of veteran teacher. If such pressure is used differentially by parents of more academically able students, or by parents more likely to be able to provide for their children a good educational environment, then disentangling the effects of differential assignment from actual differences in teaching effectiveness would be more difficult (see Clotfelter, Ladd, & Vigdor, 2006, for a demonstration of this effect).The inability to randomly assign students to teachers means, again, that researchers must identify, reliably measure, and take into account student variables that are associated with achievement. In this regard we are somewhat fortunate that most school districts routinely test students at certain grade levels. This means that the potential for a relatively rich (if incomplete) description of student prior academic achievement exists, and having good baseline measurement (i.e., reliably measured and well correlated with the outcome) is key to reducing preexisting differences (Cook, Shadish, & Wong, 2008) Ideally students would be tested very early in the school year and again very late in the school year, and student growth from the pretest to the posttest would be the key outcome (i.e., a rigorous value-added approach). Unfortunately, most school districts only test students at the end of the school year. Due to the likelihood that socially disadvantaged students lose academic skills over the summer break more dramatically than non-disadvantaged

students (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996), it seems plausible that failing to take this into account will bias estimates of student growth attributable to teachers.

Further, teachers are but one of several sources of influence on student achievement. Other largely independent factors related to the student (e.g., motivation, academic ability), family (e.g., socioeconomic status, parental involvement), and school (e.g., climate, quality of the principal) also have the potential to play important roles in student achievement. As such, the proportion of variability in student scores that is attributable to teachers is likely to be small (see the effect sizes reviewed earlier), and the proportion of variability in student scores attributable to teacher characteristics knowable at the time of hire even smaller, thus making such effects more difficult to detect and studies of those effects more sensitive to errors and biases.

A final concern, particularly related to our goal of identifying characteristics of teachers knowable at the time of hire that predict student achievement, is that it is very difficult (and often impossible) to obtain data on teacher effectiveness from candidates who were not hired by the district. Some potential candidates might be employed by other districts – and their effectiveness is theoretically knowable even if logistically the information is difficult to obtain – but others may never be hired to work as teachers. If the hiring process generally works as intended (so that better teachers have a higher probability of being hired) then this will introduce an artifact (range restriction) that

will serve to attenuate observed relations. In other words, characteristics of teachers that actually do predict student achievement will be more difficult to detect.

Taken together, these concerns suggest that – regardless of the specific aspects of teachers that are being studied – the most appropriate role for analyses linking teachers to student achievement data is in identifying teachers who might benefit from additional professional development. In essentially all cases the data will simply be too "muddy" to be useful as an important or high stakes consideration in, for example, termination or tenure decisions. With this caveat in mind, we turn next to a description of the kinds of information on applicants that is collected by typical districts, focusing on our experiences with one district.

Data Collected by the Jefferson County Public Schools

Serving a community of over 700,000 residents and with over 97,000 students, Jefferson County Public Schools (JCPS) is a large, urban district in north-central Kentucky. The district employs approximately 6,500 teachers, and hires approximately 550 annually (from an applicant pool of around 2,000). Applications are not spread equally among positions, and some are harder to staff than others. For example, JCPS may receive 6-7 applications for every elementary position available, but only 1.2 for every middle school mathematics position available. Like many districts, JCPS's application process requires that applicants provide a resume, official transcripts, teaching certificates, references, and previous teaching evaluations. JCPS also requires

that applicants complete the Teacher Disposition Survey (TDS)[1]. Applicants who meet eligibility criteria have the opportunity for site-based interviews, and hiring decisions are ultimately made at the school level.

Some of the data collected by JCPS exists in electronic format that is easily manipulated through database applications. However, some information that might be predictive of student achievement does not exist in a way that makes it readily usable for data analysis. Examples include the institutions attended and candidate grade point average. Further, other information that might have predictive value is not currently collected by the district. Examples include scores on standardized tests such as the SAT, ACT, and the GRE, the number of times that the licensing exam was taken, and the actual scores obtained on that exam, in addition to broader indices of teacher traits that are more difficult to measure easily. We return to this point in the *Implications and Future Directions* section.

Finally, it must be noted that while JCPS does link teacher information with student achievement data, the database system is currently not able to fully support a rigorous analysis that attempts to take into account the many contextual factors that might be mistaken for a relationship between teacher characteristics and student achievement. In particular, the system needs to be able to support a multi-level model that includes school-level, classroom-level, and student-level variables (Doran, 2003).

[1] The TDS was developed locally and first administered for teachers hired in the 2009-2010 school year. Previously, JCPS had used a structured interview conducted by experienced teachers trained in the interview method.

Much of this information is already collected (either internally or externally, such as data available via the Common Core of Data from the National Center for Education Statistics) or is relatively easily collectable, so with minimal investment JCPS will be well-positioned to link student achievement data with teacher data in a relatively objective and believable manner.

Perhaps a more important and difficult task is identifying teacher characteristics that are worth tracking – that is, those characteristics that, alone or in combination, might be related to student outcomes. To help with this process, we conducted a systematic review of studies that have examined the relation between teacher characteristics known at the time of hire and student achievement (and other variables related to teacher quality, such as principal evaluations). Systematic reviews are meant to overcome many of the limitations of traditional narrative reviews of research (details can be found in texts such as Cooper, Hedges, & Valentine, 2009, and Lipsey & Wilson, 2001). In particular, in a systematic review researchers set out with the goal of uncovering all studies relevant to the research question, not just those that are easy to find or are already known to the research. In part this goal serves to overcome the problem of publication bias (the well-known tendency for the published portion of a literature to over-represent statistically significant findings – all else being equal, this means that published studies tend to overestimate true effects). In addition, researchers conducting a systematic review use a pre-defined and reproducible process for finding

literature and extracting relevant data from the studies, and also take steps to assess and maximize the reliability of the retrieval and coding processes. Finally, many systematic reviews culminate in a meta-analysis, a formal statistical analysis that is meant to overcome the weaknesses of the more subjective and/or statistically weaker techniques often employed in by those conducting literature reviews (see Valentine et al. in press for a description of these techniques and their limitations). We were unable to locate enough literature to support such an analysis at this time, for even though we uncovered a reasonable number of studies there was not enough overlap in terms of the teacher characteristics that were analyzed in the studies. As such, below we discuss the potential synthesis methods that could be used once more studies are available.

Methods of the Systematic Review

Literature Search Procedures

We used two complimentary procedures to locate studies of the relation between teacher characteristics and student outcomes (as well as other outcomes, such as principal ratings and teacher retention). First, we carried out an electronic search in several academic databases (including PsychInfo, ERIC, and Proquest Digital Dissertations). Briefly, we used terms that were designed to find studies of characteristics (such as background or personality or GPA) of effective (quality or effective) teachers (teacher or teaching) knowable at the time of hire (hire or hiring). This process resulted in the identification of 784 potentially eligible studies. Based on

their titles and abstracts, a subset of potentially eligible studies was judged for eligibility by either two or three researchers. Because judging eligibility involved only four low inference judgments, agreement was quite high, and the remaining studies were examined by only one researcher. This process resulted in 124 studies that were potentially eligible; the full text of these was obtained, and eligibility was assessed again. This resulted in 10 studies that met inclusion criteria.

In addition, we examined the reference lists of the 10 studies that met inclusion criteria (as well as several other studies that did not meet inclusion criteria), looking specifically for other studies that might meet inclusion criteria. This resulted in the identification of four additional studies, bringing the total of eligible studies to 14.

Two final comments relevant to the inclusion of studies are warranted. First, our stated interest is in examining characteristics of teachers that are knowable at the time of hire and how these relate to a variety of outcomes of interest to the public and educational administrators. As such, we focused our efforts on locating studies that involved new (as opposed to experienced) teachers. The practical implication of this choice is that many studies of the impact of certification status (e.g., regularly vs. alternate certification) on student outcomes are not included in this review. We did however include studies that examined characteristics of teachers knowable at the time of hire, e.g., SAT scores, even if the teachers themselves were not new to the profession, as long as the analysis did not confound teacher experience with the variable of interest.

In addition, because the TPI has been the subject of a recent systematic review and meta-analysis (Metzger & Wu, 2008), we did not include studies of the TPI (though due to the paucity of research on the relation between scores on the TPI and student achievement identified in the Metzger and Wu paper, we would have included these had we encountered any). However, we did include evidence from other pre-hiring inventories.

Data Coding

We extracted several different types of information from studies, including characteristics of the school district and students and the number of schools, teachers, and students in the analysis. We also extracted information about the teacher characteristics that were studied (not only what they were, but how they were measured). Finally, we coded information about the dependent variables, including the specific operationalization and any information given regarding reliability and validity.

Options for Synthesis

Virtually all studies of the effect of teacher characteristics on student outcomes will use regression techniques to estimate the relations. As such, studies will report regression coefficients (usually standardized) as the effect size. The question is how to best synthesize these studies (i.e., how to arrive at summary conclusions about what the studies reveal regarding the relationships being examined). Below we discuss several strategies.

Vote Counts Based on Statistical Significance

Perhaps the most common approach is known as "vote counting." To conduct a vote count, the scholar gathers a set of studies that pertain to the research question and then determines what each study "says" about that relationship. In its most typical form, this determination is made on the basis of the statistical significance of the results. So, if Study A finds a statistically significant positive effect, the scholar casts a "Positive" vote for the relationship. If Study B finds a statistically significant negative effect, the scholar casts a "Negative" vote for the relationship. If Study C did not find a statistically significant relationship, the scholar casts an "Undecided" vote; note that statistical significance is the only criterion. Other information, notably the direction of the relationship and the magnitude of the effect, is not taken into account (this turns out to be an important point; see *Vote Counts Based on Observed Direction*, below). This process continues until the scholar has polled all of the studies, at which point the category with the most votes "wins".

Vote counting is based on a pervasive but incorrect belief about the interpretation of the probability values arising from tests of statistical significance. When conducting a test of statistical significance for the relationship between two variables, the probability value is the chance of observing a relationship at least as large as the one observed, given a true null hypothesis (Cohen, 1994). Another way to think about the probability value is that it represents the confidence with which we can state

that we have correctly identified the direction of the relationship. The relation between the probability values in one study and the likelihood of successful replication in even an exactly replicated second study is therefore not straightforward. If a study rejects the null hypothesis at $p = .05$, for example, that does not mean that the next study (even if it is an exact replication) has a 95% chance of rejecting the null hypothesis (if the population and sample effect sizes are very similar the probability is actually closer to 50%; see Greenwald, Gonzalez, Harris, & Guthrie, 1996; Valentine, Pigott, & Rothstein, 2010).

The limitations of vote counting based on statistical significance are therefore well-known: The majority of studies must have statistically significant results in order for the claim to be made that a relationship "is real." Unfortunately, in most circumstances when using vote counting it is unacceptably probable that studies will *not* reach the same statistical conclusion, even if they are estimating the same population parameter (e.g., if the intervention really is effective). For example, if two independent studies are conducted with statistical power of .80 (meaning that both have an 80% chance of correctly rejecting a false null hypothesis), in only 64% of cases will both studies result in a correct rejection of the null hypothesis. If both studies are conducted with statistical power of .50, then in only 25% of cases will both studies result in a correct rejection of the null hypothesis. As such, because studies are typically not highly powered, in most current real-world contexts requiring studies to reach the

same statistical conclusion is an approach with an unacceptably high error rate (by failing to detect real intervention effects when they exist). In fact, Hedges and Olkin (1985) demonstrated the counterintuitive result that, in many situations common in educational (i.e., interventions with moderate effects investigated in studies with moderate statistical power), vote counting based on statistical significance can actually have *less* statistical power the more studies are available. As such, comparing the statistical significance of the results of studies—while intuitively appealing and simple to implement—is a seriously limited inferential technique that is not well-suited to identifying effective interventions.

Meta-Analysis

A generally better alternative to vote counts based on statistical significance is meta-analysis (literally, "after" or "beyond" analysis), which is a statistical technique for combining the results of multiple studies that address conceptually the same research question. Meta-analysis involves statistical techniques that are analogous to the analysis of variance (ANOVA) and multiple regression. Techniques like ANOVA and multiple regression are based in part on the assumption that the residual variance associated with every observation is the same (the homoscedasticity assumption). The residual variance of an effect size is related inversely to the sample size with which the effect is estimated. Therefore, a collection of studies with varying sample sizes -- which is almost always the case -- will violate the assumption of homoscedasticity. Meta-

analysis addresses the problems presented by studies with different sample sizes by weighting effect sizes. More specifically, studies are weighted in a way that gives relatively more influence to effect sizes that come from relatively larger sample sizes. These weights are then used to create a weighted mean effect size and confidence interval for the set of studies. The logic here is the same as the logic underlying weighted means in general. For example, assume a scholar is interested in finding the average grades earned by students in a particular school. One class of 15 is surveyed, and the mean GPA was 3.5. Another class of 30 was surveyed, and the mean GPA was 2.5. If you were to take the straight mean of the two classes, you would conclude that the average student has a GPA of 3.0. However, it is clear that this strategy ignores the fact that your estimates are based on different sample sizes, and the resulting mean is different from the mean that could have been computed from the original student-level data.

A weighted mean overcomes this problem. The formula for computing a weighted mean is

$$\bar{Y} = \frac{\Sigma w_i Y_i}{\Sigma w_i} \tag{1}$$

where $\bar{Y}$ is the weighted mean, w_i is the weight for each observation i, and Y_i is the mean for each observation i. To complete the example,

$$GPA = \frac{\Sigma w_i Y_i}{\Sigma w_i} = \frac{(15 \times 3.5) + (30 \times 2.5)}{15 + 30} = \frac{52.5 + 75}{45} = 2.833$$

Note that 2.833 is the same answer that would be obtained if the mean had been computed from the student-level data instead of the classroom-level data.

The formula for computing a weighted mean effect size in meta-analysis will now seem very familiar:

$$\overline{ES} = \frac{\Sigma w_i ES_i}{\Sigma w_i} \quad (2)$$

where $\overline{ES}$ is the weighted mean effect size, w_i is the weight for study i, and ES_i is the effect size for each study i. The only complexity is determining what the correct weight to use. Generally, these are based on the standard error of the effect size; these are given in most texts on meta-analysis, but conceptually it is sufficiently accurate to think of the weight as being a function of sample size, with larger studies given more weight.

Most statistics students learn, without ever being told, how to carry out ANOVA and multiple regression using what is known as a fixed effects model. A rarely-taught alternative is the random effects model. A conceptually similar choice exists for those carrying out a meta-analysis, though the random effects model is much more often utilized in meta-analysis than it is in the analysis of primary studies. In meta-analysis, researchers adopt a fixed effects model when they believe that studies are estimating a single underlying population parameter (that is, the null hypothesis is that the studies are all estimating a single population value of zero) and any variation in observed effects across studies is presumed to be a function of random sampling error (an

assumption that can be evaluated using the homogeneity test). One important implication of this choice is that the confidence intervals around a weighted mean effect size are essentially only a function of the sample size of the underlying studies: The larger the total sample size, the narrower the confidence interval.

In the random effects model, studies are not presumed to be estimating the same population parameter but instead are presumed to be drawn from a population of effect sizes (that is, the null hypothesis is that the studies are drawn from a distribution of effect sizes that have a mean of zero). Another way to conceptualize this difference is to think about the null hypotheses for the two models. In both the fixed and random effects models the mean population effect is zero, but in the random effects case the mean of a distribution is being estimated, rather than one single value (Borenstein et al., 2009).

The choice of models has important consequences. Confidence intervals computed using random effects assumptions will never be smaller than their fixed effects counterparts, and will usually be larger. Statistical power is almost always greater for the fixed effects model. Nonetheless, given that virtually all studies vary from one another in multiple identifiable and unidentifiable ways, most scholars of meta-analysis believe that the random effects model is generally the most appropriate.

Vote Counts Based on Direction Observed

The foregoing discussion suggests that the results of a series of studies on the

same research question are usually best synthesized using meta-analysis of the effect sizes (Valentine et al., in press). For studies of the relation between teacher characteristics and student outcomes, this means that meta-analysis of the standardized regression coefficients would seem to be the best analysis strategy. Unfortunately, one difficulty associated with meta-analyzing regression coefficients is that the specific models generating these differ from study to study. In other words, in all likelihood no studies will produce regression estimates from models that examined the relation between teacher characteristics and student achievement using the same constructs as control variables. Given the complexity of studies of this kind, this is not a surprise. However, it means that the regression coefficients have somewhat different meanings (and different standard errors) from study to study, and as a result standard meta-analysis of the effect sizes may not be the most appropriate data synthesis technique.

An alternative is to conduct a vote count of the *directions* of the effects observed, regardless of the statistical significance of those effects. This procedure is outlined by Hedges and Olkin (1985) and Bushman and Wang (2009). Effects are categorized as positive or negative depending on the direction of the effect observed. The idea underlying this analysis is that there is a fair amount of information in the proportion of positive to negative effects (as population effects get larger, we expect that the proportion of positive to negative effects would increase). For example, assume that several studies measure the relationship between teacher handedness and student

outcomes. Presumably, there is no relation between handedness and student achievement, so we would expect that about 50% of the studies would find a positive association, and about 50% would find a negative association. The converse is true as well: Variables that are highly positively related should yield a high proportion of positive effect sizes. As an example, Greenwald, Hedges, and Laine (1996) located 24 estimates of the relation between teacher ability and student achievement, and 21 (88%) of these were positive.

The main drawback of this technique is that it generally has much lower statistical power compared to traditional meta-analysis. As such, analyses of effects that are likely to be small are more prone to Type II errors (and the relation between teacher characteristics knowable at the time of hire and student outcomes will almost certainly be small). Relative to typical vote count methods, however, the method we chose has two notable advantages. First, methods are available that allow for studies to contribute proportionally according to their size. Because larger studies generate more precise effects, weighting studies by their sample size provides better estimates of population effects. The second advantage of this method is that it allows the researcher to estimate effect sizes and confidence intervals, something that is not possible using traditional vote count methods.

One final complexity is that studies will usually present multiple models, and as such it is not clear which model should be represented in the meta-analysis. One

suggestion is to extract the directions of the effect from two models, if available. First, the unadjusted model might be used – the one with no control variables at all – if presented. Next, the model that represented the most control over extraneous variables might be used. Often, this model will be presented as the researchers' "final" model. A vote count of the directions of the effects could then be carried out on both sets of data.

Results of the Systematic Review

Table 1 presents the results of studies that examined the relation between teacher characteristics knowable at the time of hire and student achievement, while Table 2 presents the results of studies that examined teacher characteristics and other outcomes. The tables are sorted by construct measured. Even though there were not a very large number of studies, taken together, the studies involved an impressive number of teachers (over 60,000) and students (over 2,000,000). The studies examined a variety of specific operationalizations of both the teacher characteristics and the variables related to those characteristics. The most widely studied constructs were teacher credentials and teacher knowledge, both of which were operationalized by measures that are relatively easy to obtain. Below we address each of the teacher constructs measured and the evidence linking those constructs to student achievement.

Academic Preparation

Three studies examined the relation between teacher academic preparation and student outcomes. One study compared the achievement of students taught by an

education major to students taught by non-education majors (students of education majors performed less well). Three studies examined the selectivity of the teachers' undergraduate institutions in a total of five analyses. For two of the analyses, the authors reported only that the effects were not statistically significant, so the directions of the effects could not be determined. For the other three analyses, teachers from more selective institutions were associated with students who tended to score higher on the achievement tests. Finally, one study concluded that mathematics achievement was higher in students whose teachers were mathematics or science majors, as opposed to other majors.

Cognitive Ability

Only one study examined the relation between a measure of teacher cognitive ability and student outcomes (Rockoff et al., 2008). This study suggested that teachers with higher levels of cognitive ability were associated with students who tended to score higher on the achievement tests.

Credentials

Six studies, reporting 16 analyses, examined the relation between teacher credentials and student achievement. Three of these studies focused on the obtainment of advanced degrees while the other three examined the effects of certification levels on student achievement.

Advanced degree. Three studies, reporting four analyses, examined the impact of teachers having an advanced degree. In three of four studies, results suggested that having a teacher with an advanced degree was associated with higher levels of student achievement. One study concluded the opposite.

Certification. Five studies examined the relation between certification status and student achievement. In three of the studies, students of regularly certified teachers did less well than students of non-regularly certified teachers. In one study, this pattern was reversed, and in another study, the results were split between students of regularly certified teachers doing better and doing worse relative to non-regularly certified teachers. Three studies also explicitly compared teachers who entered the profession via the Teach for America program (TFA; Teach for America, 2010), comparing the achievement of their students to that of students of regularly certified teachers (in two studies) or novice teachers (in one study; presumably not all of the novice teachers were regularly certified). In most analyses students of TFA teachers outscored students of other teachers. It should be noted that Darling-Hammond et al. (2005) found that TFA teachers who were certified had students who performed better than students of regularly certified teachers. However, TFA teachers who were either uncertified or

alternatively certified had students who performed worse than students of regularly certified teachers.[2]

Interview

One study (Rockoff et al., 2008) examined teacher scores on a commercial interview instrument. This study found that teachers who scored in the top group on the interview had students who scored higher on the achievement tests than did teachers not in the top group.

Teacher Knowledge

Seven studies, reporting 14 analyses, studied the relationship between a pre-hire measure of teacher general content and pedagogical knowledge and student achievement.

Content knowledge. Seven studies, reporting 17 different analyses, examined the relation between a pre-hire measure of teacher general content knowledge (such as scores on the SAT or GRE) and student achievement outcomes. Fourteen of the 17 analyses suggested that teachers who scored higher on a measure of knowledge were associated with students who achieved at higher levels. One effect was reported as being exactly zero. Notably, the two negative effects were both found for college grade point average, the only analyses in which this variable appeared.

Pedagogical knowledge. One study (Rockoff et al., 2008) examined pedagogical

[2] These results are not included in Table 1 because Darling-Hammond et al. did not provide an overall comparison of TFA teachers vs. regularly certified teachers.

knowledge, and found that teachers who scored higher on a measure of pedagogical knowledge were associated with students who achieved at higher levels.

Psychological Traits

One study (Rockoff et al., 2008) examined the relation between teacher psychological traits and student achievement. This study found that teacher conscientiousness, extraversion, general self-efficacy, and personal self-efficacy were all associated with higher levels of student achievement.

Non-Achievement Outcomes

Five additional studies reported on the relation between teacher background characteristics and non-achievement outcomes (see Table 2). Generally these present a mixed bag of results. Mac Iver and Vaughn (2007) suggest, for example, that in the first three years alternatively or provisionally certified teachers are more likely that regularly certified teachers to be employed in the same district, but that after that (specifically years 4-6 post hire) regularly certified teachers are more likely to be in the system. Similarly, Ostlund (2006) found that five years post hire, teachers hired with regular certifications were more likely than teachers hired without certification to still be in the system. Of course, it is unclear how much of this effect is due to the district not renewing the contracts of uncertified teachers who do not become certified, as opposed to some other process (e.g., certified teachers being more connected or committed to the profession).

Limitations, Implications, and Future Directions

Limitations

Like most research, our work is subject to limitations that deserve highlighting. First, as we discussed in the introduction, research in this area tends to exploit readily-available measures of important constructs. As such, we uncovered few serious attempts to measure psychological traits (even cognitive ability) that might be related to teaching effectiveness and also found that student achievement was usually measured by mandated state exams. This is a critical limitation that pervades research on teaching quality. For example, the relatively poor measurement of student socio-economic status means that it is harder to be confident that this variable was properly controlled in analyses. Similarly, student achievement was measured mainly by scores on highly incentivized state tests, with little attention to the possibility of score inflation in the absence of real achievement gains. As such, it is possible that this research might be identifying teachers who are good at preparing students for the content on the state tests, and not necessarily teachers who are eliciting better achievement from their students.

Further, like all systematic reviews we are also concerned about the possibility of publication bias, which refers to the tendency for studies with statistically significant results to be more likely to be published. If published studies are more likely to be located than unpublished studies, this can introduce a bias against the null hypothesis.

We attempted to combat this problem by searching for relevant studies regardless of source, and by searching sources other than peer-reviewed journals for studies. A related concern is that our results are dependent on model specification, and researchers have high degree of influence over the models that they investigate and report. If there is bias in favor of a particular point of review (say that characteristic X should be positively related to student achievement), then it is relatively easy for an unethical researcher to vary model specifications and report the model most favorable to that point of view. Because in most cases (dissertations are a notable exception), researchers in the social sciences do not prepare protocols that describe in advance their intentions for data collection and analysis, about the only way to address this issue is to obtain the data from each researcher and re-run all analyses using a standard analysis protocol. However in this case we are fortunate, in that we believe that such data exploitation is most likely to occur in an attempt to change the statistical significance of results, as opposed to the direction of the results. That is, the researcher who would like to show that characteristic X is related to achievement is most likely to vary model specification if it is close to significance and in the "right" direction; such data exploitation would likely not alter our results, since we focus on the directions of the effects and not their statistical significance.

Implications

One message that we hope has emerged from our work is that the business of

connecting student achievement to teacher background or behavior is a tricky one. This issue has attained some urgency in recent years, as calls for tying teacher pay to student performance have become increasingly louder and are a focus of the current leadership in the U.S. Department of Education. However, a number of structural considerations reveal that this is a very difficult problem that needs a serious discussion of several issues. Among the concerns are that families can (to some extent) self-select their teachers and schools, and teachers can also self-select their schools and sometimes classes (with experienced teachers having relatively flexible choices and inexperienced teachers relatively few). Additionally, current efforts aimed at assessing student achievement tend to focus on one test – often of basic skills – and single tests virtually always sample a very limited domain of the content of interest. Further, these tests are usually highly incentivized state tests, and research suggests that growth on these tests is essentially uncorrelated with growth on other tests that are not as highly incentivized (e.g., Linn, Graue, & Sanders, 1990). Coupled with other considerations (many of which are addressed in the introduction), we believe that any efforts to relate teacher characteristics to student achievement almost invariably will be accompanied by a relatively high degree of ambiguity, which suggests to us that it will generally not be appropriate to use analyses like these to inform personnel decisions for individual teachers. However, we believe that when carefully carried out and analyzed, studies relating teacher characteristics to student outcomes can help inform the hiring process.

It should also seem clear that many factors – some educational, some non-educational – affect student achievement. Taken together, non-educational factors (such as natural ability and parental socio-economic status) probably explain the lion's share of the variability in student achievement. While teachers may explain a large percentage of the variability attributable to educational factors (i.e., teachers may be the most important educational factor), the percentage of the total variability that teachers explain is likely relatively low, and the percentage of total variability explainable by teacher background characteristics (such as teacher knowledge or certification status) is likely to be very small indeed.

That said, there is likely to be some benefit associated with an emphasis on collecting, warehousing, and analyzing data about teaching candidates. Our results suggest that measures of teacher knowledge might add information valuable to those making hiring decisions. The data also suggest that candidates from the Teach for America program seem to achieve results that are at least as good as those to whom they were compared. Further, but on shakier ground due in part to the way that effects were reported, is the idea that candidates from relatively selective undergraduate institutions might bring with them characteristics (such as increased academic motivation) that are beneficial to students. Apart from these findings, our work generally suggests areas where more research is needed. Of these, we highlight cognitive ability and other psychological traits studied by Rockoff et al. (2008) as

potentially promising avenues for research, though it should be recognized that most traits (such as extraversion) should have very small relations with student achievement. In this regard, we agree with the recommendations of Rockoff et al. that a battery of trait measures will likely prove more informative about the potential effectiveness of teacher candidates.

Future Directions for Districts

Given national trends, it seems sensible for districts to build information systems that will allow them to link and track a wide variety of types of information about teacher candidates, teachers, and students. Especially given some of the more aggressive recommendations for using student achievement data (such as in tenure and termination decisions), care will need to be taken to maximize the probability that such systems yield interpretable data. Existing systems in North Carolina, Florida, New York City, among others, could serve as potential models. That said, we reemphasize our skepticism that such information systems will yield data clean enough to support decisions like tenure and termination in specific cases. More defensible applications of analyses linking student achievement to teacher behaviors and characteristics include identifying (a) candidates for increased professional development and (b) characteristics of successful teachers that might inform hiring decisions.

The foregoing suggests that districts may wish to collect much more information on teacher candidates at the application stage, an activity that would be facilitated by

more research on the characteristics of teachers that seem to be associated with greater levels of student achievement. In addition, we strongly suspect that the analysis of student achievement will need to move beyond the end of the year, state mandated tests employed in most studies. In part this concern is reflective of the need to conceptualize achievement more broadly. In addition, however, we have expressed concern throughout the paper about the validity of state tests, largely due to research suggesting that gains on state tests are not reflected in other tests (itself suggesting that the state tests are being "overtaught"). In addition, most agree that individual teachers should be assessed in terms of student growth, but given summer learning loss this suggests that districts would be better served by testing both very early and again very late in the school year (so that growth can be measured more effectively). Given the widely held perception that there is too much testing (Barton, 1999), we suggest that a re-thinking of the way testing is currently implemented in most districts may be needed to facilitate more informative data collection at national, state, and local levels (specifically, a move away from high-stakes testing toward more diagnostic testing). This study highlights some of the tradeoffs involved in trying to balance a realistic testing regimen with the need of policymakers, administrators, and parents to be able to identify the most effective teachers.

References

*References marked with an asterisk indicate studies that contributed data to the analyses

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*, pp. 95-135.

Aloe, A., & Becker, B. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher*, *38*, 612-624.

*Ayers, J. B. & Qualls, G. S. (1979). Concurrent and predictive validity of the National Teacher Examinations. *Journal of Educational Research, 73*, 86-92.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.

Bushman, B. J. & Wang, M. C. (2009). Vote-counting procedures for meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.), pp. 207-220. New York: Russell Sage Foundation.

Barton, P. E. (1999). *Too much testing of the wrong kind; too little of the right kind in k-12 education. A policy information perspective.* Princeton, NJ: Educational Testing Service. Retrieved from ERIC database. (ED430052)

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*, 205-213.

*Boyd, D., Grossman, P., Lankford, H., Loeb, S., Wyckoff, J., & National Bureau of

Economic, R. (2005). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement. NBER Working Paper No. 11844. *National Bureau of Economic Research*, Retrieved from ERIC database.

Chesebro, J. L. (2003). The effects of teacher clarity and immediacy on student learning, receiver apprehension, and affect. *Communication Education, 52*, 135-147.

*Clotfelter,C. T., Ladd, H. F., & Vigdor,J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*, 778-820.

Coleman et al., , J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington DC: U.S. Department of Health, Eduction & Welfare Office of Education. Retrieved from ERIC database. (ED012275)

Cook, T. D., Shadish, W. R., & Wong, V. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724-750.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*, 227-268.

*Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and

teacher effectiveness. *Education Policy Analysis Archives*, *13*(42). Retrieved March 8, 2010 from http://epaa.asu.edu/epaa/v13n42/.

*Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The effects of Teach for America on Students: Findings from a national evaluation*. Mathematica Policy Research, Inc. (MPR Reference Number 8792-750).

Doran, H. C. (2003). The challenges of accountability. *Educational Leadership: The Challenges of Accountability, 61* (3), 55-59.

Greenwald, R., Hedges,L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*, 361-396.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should by replicated? *Psychophysiology, 33*, 175-183.

*Goldhaber (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources, 42*, 765-794.

*Ferguson, R. F. & Ladd, H. F. (1996). How and why money matters: An analysis of Alabama schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 265-398). Washington DC: The Brookings Institution Press.

Haberman, M. (1993). Predicting the success of urban teachers (The Milwaukee trials). *Action in Teacher Education, 15*, 1-5.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hill, C. J., Bloom, H. S., Rebeck Black, A., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177.

*Hill, H. C., Rowen, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371-406.

Konstantopoulos, S. (2008). Do small classes reduce the achievement gap between low and high achievers? Evidence from Project STAR. *Elementary School Journal, 108*, 275-291.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that `Everyone is above average." *Educational Measurement: Issues and Practice, 9*, 5-14.

*Luschei, T. F. (2006). *In search of good teachers: Patterns of teacher quality in two Mexican states*. Available through Proquest Digital Dissertation database (UMI No. 3197474)

*Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*, 615-631.

*Mac Iver, M. A. & Vaughn, E. S. (2007). "But how long will they stay?" Alternative

certification and new teacher retention in an urban district. *ERS Spectrum, 25*, 33-44.

*McAlister, K. W. (2003). *Relationships among licensure test scores, perceptions of preparedness, retention and teacher certification pathways*. Available through Proquest Digital Dissertation database (UMI No. 3117201)

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Moyer-Packenham, P. S., Bolyard, J. J., Kitsantas, A. & Oh, H. (2008). The assessment of mathematics and science teacher quality. *Peabody Journal of Education, 83*, 562-591.

*Ostlund, C. N. (2006). The Teacher Perceiver Interview as a predictor of teacher retention in special education. Available through Proquest Digital Dissertation database (UMI No. 3218434)

Phillips , K. J. R. (2010). What does "highly qualified" mean for student achievement? Evaluating the relationships between teacher quality indicators and at-risk students' mathematics and reading achievement gains in first grade. *Elementary School Journal, 110*, 464-493. doi: 10.1086/651192

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes.*

Washington, DC: Economic Policy Institute.

*Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*, 247-252.

*Rockoff, J. E., Jacob, B. A., Kane, T. J., & Steiger, D. O. (2008). *Can you recognize an effective teacher when you recruit one*? National Bureau of Economic Research Working Paper No. 14485. Cambridge, MA: National Bureau of Economic Research.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*, 537-571. doi: 10.1162/edfp.2009.4.4.537

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*, 175-214.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design*. Boston; New York: Houghton Mifflin Company.

Scribner & Akiba, 2010

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.

Teven, J. J. (2007). Teacher caring and classroom behavior: Relationships with student affect and perceptions of teacher competence and trustworthiness. Communication Quarterly, 55, 433-450. doi: 10.1080/01463370701658077

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power in meta-analysis. *Journal of Educational and Behavioral Statistics, 35,* 215-247.

*Zigarelli, M. A. (1996). An empirical test of conclusions from effective schools research. *Journal of Educational Research, 90*, 103-110.

Table 1. Studies Examining Background Characteristics and Student Achievement

Author(s) (Year)	Level	Schools	Teachers	Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
Clotfelter et al. (2006)	Elementary		3,824	≈ 60,000	Academic Preparation	Selectivity of Undergraduate Institution	State math exam	+	Students of teachers who attended more selective undergraduate institutions scored higher
Clotfelter et al. (2006)	Elementary		3,824	≈ 60,000	Academic Preparation	Selectivity of Undergraduate Institution	State reading exam	+	Students of teachers who attended more selective undergraduate institutions scored higher
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Academic Preparation	Selectivity of Undergraduate Institution	State math exam	?	Was not statistically significant, no indication given of the direction of the effect
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Academic Preparation	Selectivity of Undergraduate Institution	State English/language arts exam	?	Was not statistically significant, no indication given of the direction of the effect
Rockoff et al. (2008)	Elementary	988	4,877		Academic Preparation	Barrons Rank of Undergraduate Institution	State math exam	+	Students of teachers who attended more selective institutions outscored students of teachers from less prestigious institutions.
Rockoff et al. (2008)	Elementary	988	4,877		Academic Preparation	Education major vs. not	State math exam	-	Students of teachers who were education majors scored lower than students of teachers with other majors
Rockoff et al. (2008)	Elementary	988	4,877		Academic Preparation	Math or Science Major vs. Other	State math exam	+	Students of teachers who majored in math or science scored higher than students of teachers with other majors
Rockoff et al. (2008)	Elementary	988	4,877		Cognitive Ability	Raven's Progressive	State math exam	+	Students with a teacher who scored higher on

Author(s) (Year)	Level	Schools	Teachers	Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
						Matricies			cognitive ability scored higher
Boyd et al. (2006)	Elementary -Middle			≈ 2,000,000	Credentials	Teach for America vs. regularly certified teachers	State math exam	+	Students of teachers from the TFA program outscored students of teachers who were regularly certified
Boyd et al. (2006)	Elementary -Middle			≈ 2,000,000	Credentials	Teach for America vs. regularly certified teachers	State English/language arts exam	-	Students of teachers who were regularly certified outscored students of teachers from the TFA program
Darling-Hammond et al. (2005)	Elementary		15,334	271,015	Credentials	Certified vs. Alternatively certified teachers	State math exam	+	Students of teachers who were regularly certified outscored students of teachers who were alternatively certified
Darling-Hammond et al. (2005)	Elementary		15,334	271,015	Credentials	Certified vs. Alternatively certified teachers	SAT-9 math	+	Students of teachers who were regularly certified outscored students of teachers who were alternatively certified
Darling-Hammond et al. (2005)	Elementary		15,334	271,015	Credentials	Certified vs. Alternatively certified teachers	Aprenda math	+	Students of teachers who were regularly certified outscored students of teachers who were alternatively certified
Darling-Hammond et al. (2005)	Elementary		15,334	271,015	Credentials	Certified vs. Alternatively certified teachers	State reading exam	+	Students of teachers who were regularly certified outscored students of teachers who were alternatively certified
Darling-Hammond et al. (2005)	Elementary		15,334	271,015	Credentials	Certified vs. Alternatively certified teachers	SAT-9 reading	+	Students of teachers who were regularly certified outscored students of teachers who were alternatively certified

Author(s) (Year)	Level	Schools	Teachers	Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
Darling-Hammond et al. (2005)	Elementary		15,334	271,015	Credentials	Certified vs. Alternatively certified teachers	Aprenda reading	+	Students of teachers who were regularly certified outscored students of teachers who were alternatively certified
Decker et al. (2004)	Elementary		57	1,893	Credentials	Teach for America vs. novice teachers	ITBS Reading	+	Students of teachers from TFA outscored students of other novice teachers
Decker et al. (2004)	Elementary		57	1,893	Credentials	Teach for America vs. novice teachers	ITBS Math	+	Students of teachers from TFA outscored students of other novice teachers
Hill et al. (2005)	Elementary	115	669	2,963	Credentials	Certified vs. not (provisional + emergency?)	Standardized math test	+	Students of certified teachers scored better than students of non-certified teachers
Hill et al. (2005)	Elementary	115	669	2,963	Credentials	Certified vs. not (provisional + emergency?)	Standardized math test	-	Students of certified teachers scored worse than students of non
Hill et al. (2005)	Elementary	115	669	2,963	Credentials	Certified vs. not (provisional + emergency?)	Standardized math test	+	Students of certified teachers scored better than students of non
Hill et al. (2005)	Elementary	115	669	2,963	Credentials	Certified vs. not (provisional + emergency?)	Standardized math test	-	Students of certified teachers scored worse than students of non
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Credentials	Teach for American vs. Regularly Certified Teachers	State math exam	+	Students with a TFA teacher outscored students with a regularly certified teacher
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Credentials	Teach for American vs. Regularly Certified Teachers	State reading exam	+	Students with a TFA teacher outscored students with a regularly certified teacher
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Credentials	Uncertified vs. Regularly Certified Teachers	State math exam	+	Students with an uncertified teacher scored lower than students with a regularly

Author(s) (Year)	Level	Schools	Teachers	Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
									certified teacher
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Credentials	Uncertified vs. Regularly Certified Teachers	State reading exam	+	Students with an uncertified teacher outscored students with a regularly certified teacher
Rivkin et al. (2005)	Elementary + Middle			≈1,300,000	Credentials	Proportion of teachers with master's degree	State math exam	-	As a school's proportion of teachers with a master's degree increased, achievement decreased
Rivkin et al. (2005)	Elementary + Middle			≈1,300,000	Credentials	Proportion of teachers with master's degree	State reading exam	+	As a school's proportion of teachers with a master's degree increased, achievement increased
Rockoff et al. (2008)	Elementary	988	4,877		Credentials	Teacher has a graduate degree vs. not	State math exam	-	Students of teachers with graduate degrees scored lower than students without graduate degrees
Zigarelli (1996)	Secondary	1,100	≈ 5,000	7,402	Credentials	% of teachers with advanced degree	IRT estimated reading, math, science, history (composite)	+	Higher school proportion of teachers with advanced degree associated with better student achievement
Rockoff et al. (2008)	Elementary	988	4,877		Interview	Haberman Prescreener – top group of scores vs. others	State math exam	+	Students of teachers who scored in the top group of the Haberman had higher scores
Clotfelter et al. (2006)	Elementary		3,824	≈ 60,000	Knowledge	Scores on state licensing exam	State math exam	+	Teacher scores on licensing exam were positively related to math achievement
Clotfelter et al. (2006)	Elementary		3,824	≈ 60,000	Knowledge	Scores on state licensing exam	State reading exam	+	Teacher scores on licensing exam were

Author(s) (Year)	Level	Schools	Teachers	Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
									positively related to reading achievement
Ferguson & Ladd (1996)	Elementary	35			Knowledge	Teacher score on ACT (from official records)	Stanford Achievement Test (3rd grade) – reading subscale Basic Competency Test (4th grade) – reading subscale	+	Teacher scores on the ACT were positively associated with student achievement
Ferguson & Ladd (1996)	Elementary	35			Knowledge	Teacher score on ACT (from official records)	Stanford Achievement Test (3rd grade) – math subscale Basic Competency Test (4th grade) – math subscale	+	Teacher scores on the ACT were positively associated with student achievement
Goldhaber (2007)	Elementary		24,327	≈ 700,000	Knowledge	Scores on teacher certification exams (e.g., Praxis)	State math test	+	Students of teachers who scored in the lowest quintile scored lower
Goldhaber (2007)	Elementary		24,327	≈ 700,000	Knowledge	Scores on teacher certification exams (e.g., Praxis)	State reading test	+	Students of teachers who scored in the lowest quintile scored lower
Hill et al. (2005)	Elementary	115	669	2,963	Knowledge	Math pedagogical knowledge	Standardized math test	+	Students of teachers who scored higher on a measure of math pedagogical knowledge scored higher
Hill et al. (2005)	Elementary	115	669	2,963	Knowledge	Math pedagogical knowledge	Standardized math test	+	Students of teachers who scored higher on a measure of math pedagogical knowledge scored higher
Hill et al. (2005)	Elementary	115	669	2,963	Knowledge	Math pedagogical knowledge	Standardized math test	+	Students of teachers who scored higher on a measure of math pedagogical knowledge

Author(s) (Year)	Level	Schools	Teachers	Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
									scored higher
Hill et al. (2005)	Elementary	115	669	2,963	Knowledge	Math pedagogical knowledge	Standardized math test	+	Students of teachers who scored higher on a measure of math pedagogical knowledge scored higher
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Knowledge	SAT Score	State math exam	+	Teacher scores on the SAT were positively associated with student test scores
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Knowledge	SAT Score	State reading exam	0	Teacher scores on the SAT were not at all associated with student test scores
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Knowledge	College GPA	State math exam	-	Teacher scores on the SAT were negatively associated with student test scores
Kane et al. (2008)	Elementary		≈ 50,000	≈ 1,400,000	Knowledge	College GPA	State reading exam	-	Teacher scores on the SAT were negatively associated with student test scores
Luschei (2006)	Elementary	629	3,981	130,291	Knowledge	Score on a test of subject matter and pedagogical knowledge, as well as legal and administrative issues	Percent correct on state-administered end of grade test	+	Higher teacher scores on test associated with better student achievement
Rockoff et al. (2008)	Elementary	988	4,877		Pedagogical Knowledge	Content knowledge test (Hill, 2006)	State math exam	+	Students of teachers with higher scores on the content knowledge test scored higher
Rockoff et al. (2008)	Elementary	988	4,877		Traits	Conscientiousness	State math exam	+	Students of teachers with higher scores on Conscientiousness scored

Author(s) (Year)	Level	Schools	Teachers	Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
									higher
Rockoff et al. (2008)	Elementary	988	4,877		Traits	Extraversion	State math exam	+	Students of teachers with higher scores on Extraversion scored higher
Rockoff et al. (2008)	Elementary	988	4,877		Traits	Self-efficacy (general)	State math exam	+	Students of teachers with higher scores on general self-efficacy scored higher
Rockoff et al. (2008)	Elementary	988	4,877		Traits	Self-efficacy (personal)	State math exam	+	Students of teachers with higher scores on personal self-efficacy scored higher

Notes. Luschei (2006) studied teachers in Mexico. The Kane et al. (2008) analyses of SAT scores and GPA were limited to a subset of teachers who were classified as Teaching Fellows. Darling-Hammond et al. (2005) contained several more analyses, all generally suggesting that regularly certified teachers outperformed teachers who were not regularly certified.

Table 2. Studies Examining Background Characteristics and Non-Achievement Outcomes

Author(s) (Year)	Level	Number of Schools	Number of Teachers	Number of Students	Teacher Characteristic - Construct	Teacher Characteristic – Specific Variable	Dependent Variable – Specific Operationalization	Direction of Effect	Interpretation
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Principal rated competency	+	Teachers scoring higher on the NTE were rated higher by principals
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Principal rated relations with students	+	Teachers scoring higher on the NTE were rated higher by principals
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Principal rated appropriateness of assignments	+	Teachers scoring higher on the NTE were rated higher by principals
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Principal rated overall effectiveness	+	Teachers scoring higher on the NTE were rated higher by principals
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Observer rated creativity	-	Teachers scoring higher on the NTE were rated lower by observers
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Observer rated dynamism	-	Teachers scoring higher on the NTE were rated lower by observers
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Observer rated demeanor	-	Teachers scoring higher on the NTE were rated lower by observers
Ayers & Qualls (1979)	Elementary + Secondary		148		Knowledge	National Teacher Exam	Observer rated warmth and acceptance	+	Teachers scoring higher on the NTE were rated higher by observers
Mac Iver & Vaughn (2007)			4,302		Credentials	Alternatively and provisionally vs. Regularly certified teachers	Teacher retention	-	One year after hiring, alternatively or provisionally certified teachers more likely to be in the system
Mac Iver & Vaughn (2007)			2,710		Credentials	Alternatively and provisionally vs. Regularly certified	Teacher retention	-	Two years after hiring, alternatively or provisionally certified teachers more likely to be in the system

						teachers			
Mac Iver & Vaughn (2007)			1,805		Credentials	Alternatively and provisionally vs. Regularly certified teachers	Teacher retention	-	Three years after hiring, alternatively or provisionally certified teachers more likely to be in the system
Mac Iver & Vaughn (2007)			1,069		Credentials	Alternatively and provisionally vs. Regularly certified teachers	Teacher retention	+	Four years after hiring, alternatively or provisionally certified teachers less likely to be in the system
Mac Iver & Vaughn (2007)				627	Credentials	Alternatively and provisionally vs. Regularly certified teachers	Teacher retention	+	Five years after hiring, alternatively or provisionally certified teachers less likely to be in the system
Mac Iver & Vaughn (2007)				233	Credentials	Alternatively and provisionally vs. Regularly certified teachers	Teacher retention	+	Six years after hiring, alternatively or provisionally certified teachers less likely to be in the system
McAlister (2003)		90	306		Credentials	Regularly vs. Alternatively Certified	Self-reported intentions to stay in teaching profession	-	Regularly certified teachers more likely to report intentions to leave profession
Ostlund (2006)			339		Credentials	Certified vs. not	Retention (5 years post hire)	+	Certified teachers more likely to be employed in district five years later
Rockoff et al. (2008)					Hiring Interview	Haberman Prescreener – top group vs. others	Teacher absences	+	Top group of scores reported fewer absences
Rockoff et al. (2008)					Hiring Interview	Haberman Prescreener – top group vs. others	Teacher retention	+	Top group of scorers more likely to be employed by district