Review of Educational Research
Spring 2003, Vol. 73, No. 1, pp. 89–122

Teacher Characteristics and Student Achievement Gains: A Review

Andrew J. Wayne
SRI International
Peter Youngs
Stanford University

A large body of studies exists that examines the relationship between student achievement gains and the characteristics of teachers. To help policymakers and researchers use and build on this body of studies, this article reviews the studies systematically and synthesizes their results with deliberate consideration of each study's qualities. Determinate relationships are described for four categories of teacher characteristics: college ratings, test scores, degrees and coursework, and certification status. The review details the implications of these relationships in light of study limitations and proposes directions for future research.

KEYWORDS: academic achievement, alternative teacher certification, educational quality, outcomes of education, performance factors, teacher characteristics, teacher effectiveness.

For policymakers and researchers looking for ways to improve K–12 education, one enduring approach has been to focus on teachers. Teachers are the system's principal resource. Their salaries occupy the largest share of K–12 education budgets. And both intuition and empirical research tell us that the achievement of schoolchildren depends substantially on the teachers they are assigned.[1] Recently, issues related to teacher supply have captured national attention as a result of concerns about the aging of the teaching force and the need for new teachers. A variety of researchers, policymakers, and national organizations have been scrutinizing the issue since the landmark report published by the National Commission for Teaching and America's Future (1996). More recently, a National Research Council panel investigated teacher quality and assessment (Mitchell et al., 2001), and U.S. Secretary of Education Rod Paige issued a high-profile report on teacher quality (U.S. Department of Education, 2002).

Seeking results that might inform policymakers' thinking, researchers have undertaken studies and written syntheses that focus on different aspects of teacher policy. Some focus purely on teacher *quantity,* asking, for instance, how many teachers will be needed and how many leave the profession each year (e.g., Hussar, 1999). The present review targets teacher *quality.* A large body of studies exists that examines the characteristics of effective teachers. In their attempts to draw on this

body of studies, researchers and analysts have arrived at markedly different interpretations, perhaps because of the difficulty of systematic review.[2] The present effort attempts to use more systematic methods to make the results of this research on teacher characteristics available to policymakers and researchers.

Findings about the relationship between teacher characteristics and student achievement gains are very applicable to discussions of teacher policy. States generally specify requirements for teachers in terms of degrees, coursework, and test scores.[3] Another way that knowledge about the relationship between teacher characteristics and student achievement gains figures into policy discussions is in the identification of trends in teacher quality or in the identification of problems to be solved. Low-income students may have fewer teachers with certain characteristics, for instance (see, e.g., Ingersoll, 1996; Wayne, 2002). The knowledge contained in the present body of student achievement studies is therefore especially important.

It is important to note that, as reflected in state policy, conceptions of effective teaching in the United States have changed significantly over time. For most of the 20th century, candidates were eligible for certification as long they completed a state-approved teacher preparation program. States influenced the nature of teacher preparation by prescribing coursework in subject matter and/or education and by establishing student teaching requirements. But state policies rarely promoted specific approaches to instructional practice. This changed in the 1980s as several states implemented performance assessments for use with beginning teachers. Many of these assessments were based on process-product research on teaching and focused on a uniform set of teaching behaviors regardless of the content area or grade level taught.

Over the past 10 years, many states have adopted standards from the Interstate New Teacher Assessment and Support Consortium and the National Council for Accreditation of Teacher Education for use in making decisions about licensure and accreditation. These standards reflect conceptions of teaching that differ significantly from those previously embedded in state policies. In particular, these conceptions emphasize the context-specific nature of teaching and the need for teachers to integrate knowledge of subject matter, students, and context in making instructional decisions, engaging students in active learning, and reflecting on practice (Youngs, Odden, & Porter, in press).

To organize this review, we first describe our review methods, which include our criteria for including studies in the review, the interpretation of individual studies, and our approach to synthesizing the results obtained from the various studies. The next section then uses these methods to review the research on whether teachers with certain characteristics have a greater impact on their students' standardized test scores than teachers without these characteristics. It discusses findings and implications for four categories of teacher characteristics: ratings of teachers' colleges, teachers' test scores, teachers' degrees and coursework, and teachers' certification status. The concluding section discusses general implications for policymakers and researchers.

Methods

Our objective in this review was to create a clear interpretation of the research for policymakers and researchers interested in the relationship between teacher charac-

teristics and student achievement gains. That objective informed our methods in several ways. This section discusses the following features of our methods: the scope of the review, the interpretation of individual studies, and the synthesis of results.

Scope of the Review

Studies meeting the criteria for inclusion numbered only 21 at the time of this review. Our search began with an examination of electronic databases and existing reviews. We used the following meta-analyses and reviews: Darling-Hammond (1999a); Greenwald, Hedges, and Laine (1996); Hanushek (1997); and Wilson, Floden, and Ferrini-Mundy (2001). In addition, three electronic databases were searched for the years 1975–2002: (a) the ERIC database, which indexes journals and technical resources from *Resources in Education* and *Current Index to Journals in Education;* (b) PsychLit, which houses the American Psychological Association's *Psychological Abstracts;* and (c) EconLit, which corresponds to the American Economic Association's *Journal of Economic Literature* and the *Index of Economic Articles.* These databases were searched using the following terms: (a) teach* assess*, (b) teach* certificat*, (c) teach* characteristics, (d) teach* licens*, (e) teach* qual*, and (f) teach* test*.

Although we included most material that appears in print, which includes some books and other materials that were not peer reviewed, for practical reasons we excluded conference papers and dissertations.[4] Having identified the universe of studies typically treated in reviews of student achievement studies, we then applied a set of study design criteria. The remainder of this section describes those four criteria.

First, it should already be clear that this review focuses on studies that observe teachers' characteristics as well as the standardized test scores of these teachers' students. It is also important to consider the relationships between teacher characteristics and other student outcomes such as graduation rates, attendance at postsecondary institutions, and the acquisition of knowledge and skills not easily measured by standardized tests. In the 1990s, for example, Kentucky, Maryland, Vermont, and other states developed innovative assessment systems that measured student achievement in relation to performance standards and featured assessment methods other than multiple-choice questions. At the time of this review, however, there were no studies that both (a) included such student outcomes and (b) met the remaining criteria for inclusion in the review.

Second, the review limits its scope to the achievement of students in the United States, since the review aims to inform researchers and policymakers in the United States. We are aware of only one study excluded by this criterion (Hanushek, Gomes-Neto, & Harbison, 1996).

The two remaining criteria are intended to keep the focus on findings that are compelling as opposed to being merely suggestive. These two criteria are (a) that the study design accounts for prior achievement, which excludes numerous studies, and (b) that the design accounts for students' socioeconomic status. The latter criterion excludes only a few of the remaining studies.[5]

Accounting for both prior achievement and socioeconomic status makes a study's findings more compelling because the question "Do students learn more from teachers with this characteristic?" pertains to a causal relationship. There are many studies that examine end-of-year student test scores and teacher qualifications. But in order

to attribute any observed student achievement differences to teacher characteristics, one must rule out alternative explanations. Randomized assignment to treatment and control groups can rule out alternative explanations, but such designs have rarely been used in research on teacher characteristics. The designs that have been used and have proven sufficiently convincing are nonrandomized (quasi-)experimental designs. To reject alternative explanations absent randomization, these designs employ a theory of the determinants of student achievement. Refinement efforts persist (see, e.g., Grissmer & Flanagan, 2000; Ludwig & Bassi, 1999; Rowan, 2002; Turner, 2000), but researchers generally agree that (a) prior achievement test scores and (b) student background characteristics are required in order to minimally reduce the potential for alternative explanations of student achievement differences (see generally Ehrenberg et al., 2001; Ferguson, 1996; Rowan, 2002). Such student achievement models are called "value-added" because they assume that teachers add to students' progress during the period between pretest and posttest.

Although this approach to establishing the existence of a causal relationship clearly calls for longitudinal data, a few studies successfully exploit cross-sectional data by subtracting the achievement of a group of students from one grade level from that of a group of students in the same school at the same time but at a higher grade level. These studies are regarded as convincing here and are therefore included, although studies that use longitudinal data clearly face fewer threats to validity.

In summary, the four criteria for inclusion in this review are as follows:

1. The data collected address teachers' characteristics as well as the standardized test scores of the teachers' students.
2. The data were collected in the United States.
3. The design accounts for prior achievement.
4. The design accounts for students' socioeconomic status.

Interpretation of Individual Studies

Another aspect of our review methods that helps us create a clear understanding of the findings from this body of research involves our interpretation of findings from individual studies. This section identifies several technical considerations associated with the interpretation of individual studies and explains how these considerations affected our translation of study results.

As noted above, the studies included in this review all use a theory of the determinants of student achievement. That theory takes the form of a student achievement equation. The data collection feeds that equation: Data must include student background characteristics and pretest scores, in addition to posttest scores and teacher quality information. Usually, the data include additional school- or program-level factors that could plausibly affect student achievement (e.g., class size).

The student achievement equation always proves at least partly correct. That is, its variables jointly explain some proportion of the variation in student achievement. Statistical methods then enable us to isolate the contributions of individual variables, using the student achievement model to rule out the observable influences known to be at work.

The statistical test to determine whether a particular teacher qualification contributes to student achievement poses the research question in the negative (the null hypothesis): Could the observed student achievement patterns have occurred *even if*

the teacher qualification being studied had *not* influenced student achievement? A "no" answer (rejecting the null hypothesis) translates as a positive relationship. It means that students learned more from teachers with that particular qualification, after controlling for the other variables.

But it is important to realize that a "yes" answer (failing to reject the null hypothesis) does not rule out a relationship. Perhaps the study's sample size was too small—or the measurement error too great—to provide statistical confirmation. These problems prevent rejection of the possibility that differences occurred randomly, even when the qualification being studied does influence student achievement.

Another reason statistical methods might not detect an existing relationship is that the teacher quality variable might strongly correlate with the model's variables for student characteristics, which also influence student achievement (see, e.g., Dewey, Husted, & Kenny, 2000). This phenomenon, termed multicollinearity, makes it more difficult for statistical methods to discern relationships.

For these reasons, when translating results from individual studies, we assume that studies may establish that an observed teacher quality indicator matters but cannot convincingly show that an observed teacher quality indicator does not matter. One could show that an indicator was unimportant by using large samples, keeping measurement error low, and establishing that multicollinearity was not a problem. But no such examples appear in the literature.

When statistical methods seem to establish that a particular quality indicator influences student achievement, readers still must draw conclusions cautiously. Theory generates alternative explanations that statistical methods must reject, so a positive finding is only as strong as the theory undergirding the analysis. If the theory is incomplete—or data on the plausible determinants of student achievement are incomplete—the untheorized or unavailable determinants of student achievement could potentially correlate with the teacher quality variable (i.e., correlation between the error term and the teacher quality variable). Thus, student achievement differences that appear connected to teacher qualifications might in truth originate in omitted variables. Bias due to omitted variables can influence effect size estimates both positively and negatively (see Ludwig & Bassi, 1999), so reported estimates of effect size are usually regarded as biased.

Of particular importance for the interpretation of positive findings are omitted teacher quality variables. Suppose a data set contains only one teacher quality variable: whether the teacher has a master's degree. Statistical methods may show that the master's degree matters, but another plausible explanation for that finding exists. It is likely that teachers with a master's degree have more teaching experience. Thus, what appears attributable to a master's degree may instead be attributable to experience. Studies that assess multiple teacher characteristics simultaneously are therefore more readily interpreted.

A related concern is lagged effects. Value-added models usually assume that a student's prior test score captures the effects of all previous educational experiences. However, it is likely that the effects of educational experiences may not manifest themselves immediately. To the extent that the sources of lagged effects are in any way connected to the likelihood that students will have certain teachers, estimates of the effect of having teachers with particular characteristics may be biased.

One last concern in interpreting positive findings is termed aggregation bias. In his earlier review of student achievement studies, Hanushek (1997) showed that

studies that aggregate data into larger units of analysis (e.g., school- or district-level averages rather than individual student data) are more likely to show positive effects. Thus, positive findings that emerge from studies that use higher levels of aggregation may result from aggregation bias. This review gives less weight to studies with higher levels of aggregation and excludes studies that aggregate to the state level. The only study excluded by this criterion had no determinate results regarding the influence of teacher characteristics.[6]

In summary, while quasi-experimental studies of student achievement deserve attention, interpretations must be guarded. First, researchers have not used these methods to prove that any particular teacher characteristic does *not* matter. Accordingly, we treat indeterminate findings as just that—indeterminate. Second, when studies do find relationships, we weight studies according to design features used to guard against spurious findings. Finally, it is clear that reported effect sizes are unlikely to be unbiased, and we cannot predict whether actual effects are larger or smaller. Accordingly, this review does not attempt to discuss observed effect sizes.

Synthesis of Results

Faced with hundreds of student achievement studies, researchers have employed two basic approaches to the synthesis of results. One approach draws conclusions by tallying the studies' results; thus, for each school input of interest (e.g., class size), the reviewer computes the numbers of positive, negative, and insignificant results that have appeared in the literature. The second approach to synthesis is formal meta-analysis, which attempts to pool the statistical power of many small studies that focus on the same school input.

Arguably, however, neither of these two approaches fits the present task, which involves a relatively small number of studies. The strength of the standard approaches is that they enable researchers to quickly summarize results from large numbers of studies (hundreds) and for multiple types of school inputs (not only teacher characteristics).

The approach used here avoids two pitfalls shared by the standard approaches. First, synthesis requires some judgments about the strengths and weaknesses of individual studies. The subsection above explained that determinate findings are sometimes not reliable; the more trustworthy findings are those arising from studies that analyze individual students and their teachers and include a thorough set of controls for other potential determinants of student achievement. Thus, some studies deserve more weight than others and may even refute the findings of others.

An additional limitation of the standard approaches is that they necessarily abstract details that are important to users of the findings. For example, reviews commonly lump together all studies that involve a teacher test score. But policymakers face decisions about what to test (e.g., basic academic skills, subject matter knowledge) and who should be tested (e.g., only secondary teachers or all teachers).

For these reasons, the present review employs an alternative approach to synthesis. It begins by considering groups of studies that focus on a particular teacher characteristic. It then explicitly describes each study, focusing on the features that affect the inferences that can be drawn. Finally, with due consideration of those quality features, it renders a joint interpretation.

For those accustomed to the standard approaches, this method may seem less scientific. But it is arguably better suited to the treatment of smaller numbers of studies

and for clearly communicating results to policymakers and the research community. Moreover, it maintains a solidly scientific character. The following three features of scientific inquiry identified by King, Keohane, and Verba (1994, pp. 8–9) are all present: The approach makes clear what the evidence is; it explicitly treats uncertainty; and it can be replicated by other researchers.

A final note is necessary here to discuss a commonly recognized source of bias to which both the present approach and the standard approaches are susceptible. The typical standard in social science research is that a relationship is considered determinate only if there is less than a 1-in-20 chance that it occurred randomly. Thus, if there are 20 studies, and all are indeterminate except for one, a reviewer cannot claim to have found a relationship. Moreover, because large and statistically significant effects are considered most interesting by publication outlets, it is likely that the universe of published studies is not representative of all analyses that have been undertaken. Determinate findings with large effect sizes are more likely to reach the literature than indeterminate findings. Termed "publication bias," this phenomenon affects both effect size estimates and judgments about whether results are statistically significant. This review does not address effect size, but its conclusions about the relationships that exist are subject to publication bias.

Findings

This section presents the findings from our review. For each teacher characteristic for which evidence exists, we (a) describe all relevant studies and findings, (b) render joint interpretations, and (c) consider implications for policy and future research.

The subsections to follow correspond to five categories of teacher characteristics: teachers' college ratings, test scores, course taking and degrees, certification status, and all other characteristics. Conclusive evidence is available only for the first four. The fifth category acknowledges other characteristics for which existing evidence is not conclusive.

In describing relevant studies and findings, we present results at a level of detail intended to balance content with readability. Relatively more detail is provided for the studies that yielded determinate findings. For easy access to key technical information for each study, the Appendix provides a table with the following information for all 21 studies: (a) a description of the student achievement data and sample; (b) exact specifications of teacher quality variables; (c) a list of controls, including socioeconomic status; and (d) a list of other important study features.

Ratings of Teachers' Undergraduate Institutions

Relevant Studies and Findings

Only three research efforts have sought to determine whether students learn more from teachers who graduated from better-rated undergraduate institutions. The first, Summers and Wolfe (1975a, 1977), was a set of studies undertaken in Philadelphia during the 1970–1971 school year, using samples of students in the 6th, 8th, and 12th grades during that year. School records included scores for the 1970–1971 year and some earlier years, so the authors were able to observe each student's gains over time. The three analyses examined gains from 3rd to 6th grade, 6th to 8th grade, and 9th to 12th grade.

Whereas most studies to date had treated school inputs as school-level averages, Summers and Wolfe used data specific to each student. School records contained

information on students' current and previous teachers, including teachers' scores on the National Teacher's Examination, teachers' years of experience, and the Gourman rating of each teacher's undergraduate institution. According to Summers and Wolfe, the Gourman report utilized information from the institution's and other sources on each institution's facilities, departments, administration, faculty, services for students, alumni support, and other general areas.

The analyses of 6th graders used 627 students' Iowa Test of Basic Skills (ITBS) composite scores. The authors found that Gourman ratings associated with 6th-grade teachers were important. The authors also tried aggregating the Gourman ratings and all other school inputs into school-level averages and found no relationships—a somewhat counterintuitive result given the typical direction of aggregation bias.

The analyses of 8th-grade students also examined composite ITBS score gains and used information on students' English, math, and social studies teachers. Using a sample of 553 students, the authors found that the only determinate relationship between student achievement and Gourman ratings was a positive relationship associated with 8th-grade social studies teachers.

The analysis of 12th-grade students used 716 students' scores from the California Aptitude Test and the Comprehensive Test of Basic Skills. Focusing on reading score gains, the authors used teacher quality data on the students' English teachers only. Unlike the results for the 6th and 8th grades, no determinate relationships were found among the 12th-grade students.

Murnane and Phillips (1981) conducted the second of the three studies involving college ratings. A welfare reform experiment in Gary, Indiana, in the early 1970s provided achievement data on several hundred Black elementary school students, mostly from low-income families. When the authors linked the students to their individual teachers and controlled for several other teacher characteristics, they could not discern any relationship between students' ITBS vocabulary score gains and teachers' college ratings. The study provided no information about the college rating system used.

The only other study of whether students learn more from teachers from better-rated undergraduate institutions was conducted by Ehrenberg and Brewer (1994). The authors used the High School and Beyond (HS&B) data set, which tested a sample of 10th graders in 1980 and then retested them as 12th graders in 1982. The measure of student achievement was a composite score combining mathematics, reading, and vocabulary skills.

The original HS&B data collection did not contain teacher data, but a follow-up survey of 25 teachers at each of about 320 public schools allowed Ehrenberg and Brewer to examine whether individual students learned more when their schools' average teacher quality was higher. Teacher quality variables included the percentages of teachers with master's degrees and with 10 or more years of experience. For each school, the authors also calculated the average rating of teachers' undergraduate institutions using Barron's six-category selectivity rating system. This rating system reportedly examines the entering class's entrance examination scores and high school records, in addition to the percentage of applicants admitted.

The authors performed their analyses separately on students in different race/ethnicity categories. They found that with White students and Black students, teachers from better-rated undergraduate institutions were more effective. Findings were indeterminate for Hispanic students.

Joint Interpretation

Although implications may differ by rating system, as discussed below, the proper joint interpretation of the three studies is that some relationship exists between college ratings and student achievement gains, as noted earlier. Researchers were not always able to discern a relationship, but those relationships that were found were positive. Furthermore, the two studies showing positive relationships had different strengths. As is detailed in the Appendix, the controls used by Ehrenberg and Brewer for students' socioeconomic status were student-level controls and were extensive (parents' education, family income, family structure, and family size). Ehrenberg and Brewer also adjusted their analysis to correct for students' dropout behaviors and attempted to address bias due to omitted variables using a technique called instrumental variables.

By contrast, the strengths of Summers and Wolfe's study were that all analysis was done at the student level and that the authors controlled for a large number of schooling factors, including student attendance, peer group measures, school structure and disciplinary atmosphere, and other inputs. Also, in their analyses of gains from 6th to 8th grade and 9th to 12th grade, Summers and Wolfe controlled for a third test score, from an even earlier school year, to provide additional assurance.

Implications

Although the findings encourage researchers to further examine characteristics associated with teachers' undergraduate institutions, the implications for policymakers differ depending on which rating system is involved. In the case of Barron's selectivity ratings, the finding suggests that policymakers should encourage better screening of prospective teachers. The traditional screening tool for state policymakers has been licensure tests, not college selectivity ratings. The next section discusses licensure tests more fully.

In the case of the Gourman report ratings, the ratings apparently intended to capture institutional quality, broadly defined. Insofar as institutional quality may be at play, policymakers may wish to require that teachers hold degrees from institutions with particular quality characteristics. Third-party accreditation is the typical policy instrument by which institutional quality is assured. Researchers seeking to inform policy might therefore examine the relative effectiveness of teachers from institutions with different accreditation statuses.

Test Scores[7]

Relevant Studies and Findings

Seven sophisticated studies of student achievement have assessed the importance of teachers' scores on tests of verbal skills and other tests. For organizational purposes, this subsection describes them in three groups: (a) studies involving teacher licensure examination scores, (b) subsequent student achievement studies involving tests of teachers' verbal skills, and (c) more recent studies involving other test score measures.

Teacher licensure examination scores. Two student achievement studies have examined whether students learn more from teachers who performed better on teacher licensure examinations. The two studies employed scores from the National

Teachers Examination (NTE) and from the Texas Examination of Current Administrators and Teachers (TECAT).

The use of teacher tests in making licensure decisions was rare until the 1980s. During that decade, the number of states employing tests of verbal skills, content knowledge, and/or professional knowledge for licensure dramatically increased, to more than 40. As of 2002, 41 states used tests in one or more of these areas to make decisions about admission to teacher education programs or initial licensure.

The NTE was developed by the Educational Testing Service (ETS) in the 1940s, and by the 1970s it was being used by some states in making licensure decisions (see Haney, Madaus, & Kreitzer, 1987). As teacher testing increased in the 1980s, it became the most widely used licensing examination. Specialized teachers took the NTE Area Examination, covering both content and teaching methods specific to one of about 30 specialties. Other teachers took the NTE Common Examinations, which included the following four component tests: (a) general principles of pedagogy and psychological and social foundations of education; (b) written English expression; (c) social studies, literature, and fine arts; and (d) science and mathematics. The NTE remained the most prevalent licensing exam until ETS replaced it with Praxis in the 1990s.[8]

The one qualifying student achievement study that involved teachers' NTE scores was conducted by Summers and Wolfe (1975a, 1977) using students in Philadelphia schools in 1971, as described above.[9] Among the 627 sixth graders in their elementary school sample, students learned *less* when their teachers scored higher on the NTE Common Examinations. The junior high and senior high school samples employed teachers' NTE Area Examination scores and yielded indeterminate results.[10]

The only other study to employ scores from an actual licensure examination used the TECAT, a test that evaluated reading and writing skills, integrating content related to professional knowledge. Ferguson (1991, 1998) took advantage of TECAT scores available for Texas teachers tested in statewide teacher testing in 1986 (see Kain & Singleton, 1996). He focused his analysis on reading scores, which measured teachers' reading skills and professional knowledge.[11]

Ferguson computed mean teacher TECAT scores for each Texas school district from which data could be obtained.[12] Data on students were also aggregated to the district level. Students' reading and math scores came from the 1986 Texas Educational Assessment of Minimum Skills (TEAMS) exams, which were multiple-choice tests administered to students in the 1st, 3rd, 5th, 7th, 9th, and 11th grades.

Analyzing data from almost 900 school districts, Ferguson computed the difference for each district between the mean achievement scores of 1st, 3rd, 5th, and 7th graders in 1986 and the mean scores of 5th, 7th, 9th, and 11th graders in 1990—the same cohorts of students assuming that student migration is negligible. He found that districts where teachers had higher TECAT scores were more likely to have higher gains in student test scores in reading, especially between 3rd and 7th grades. Ferguson later reinforced these findings (see Ferguson, 1998) by showing that the gains of each district's elementary students differed from the gains of its secondary students, depending on the TECAT score differences between the district's elementary teachers and its secondary teachers.

Verbal skills. Three subsequent student achievement studies employed tests of teachers' verbal skills and used relatively old data. One reanalyzed the data from the

Equality of Educational Opportunity study (Coleman et al., 1966), a very large cross-sectional study conducted in the mid-1960s. In their reanalysis, Ehrenberg and Brewer (1995) constructed synthetic gain scores by subtracting schools' lower-grade-level achievement averages from their higher-grade-level achievement averages. Specifically, the authors subtracted the average achievement scores of 3rd-grade students in 969 elementary schools from those of 6th-grade students in the same schools and subtracted the average achievement scores of 9th-grade students in 256 secondary schools from those of 12th-grade students in the same schools.

After teachers' experience and graduate education had been controlled, teachers' scores on a short verbal facility test explained some school-to-school variation in the gain scores. Furthermore, the authors' attempt to use the technique of instrumental variables to address bias due to omitted variables did not change the findings.

The remaining two studies involving teachers' verbal skills both analyzed data from the Gary Income Maintenance Experiment, a welfare reform experiment conducted in Gary, Indiana, in the early 1970s that, as mentioned earlier, provided achievement data on several hundred Black elementary school students. The data set included a variety of data on students' teachers, including scores on a test labeled "Quick Word Test."

In their analysis, Murnane and Phillips (1981) focused on students' ITBS vocabulary scores. Controlling for several other teacher characteristics, including race, sex, experience, possession of a master's degree, and the rating of the teachers' undergraduate institution, they found no relationship between the teachers' word test scores and student achievement gains. However, when they disaggregated the students by grade level, they found that the 6th-grade students learned *less* when their teachers had higher scores on the word test. They rejected this finding as spurious "because a large number of teachers used aids in completing the test" (pp. 97–98).

Hanushek (1992) examined the same data, but his dependent variables were changes in achievement on *both* the ITBS reading and the ITBS vocabulary tests. He examined changes among 2nd- through 6th-grade students across a single grade level (e.g., achievement growth from the end of 2nd through the end of 3rd grade). Hanushek reported that teachers' performance on the word tests affected their students' reading score gains but not their vocabulary score gains. Unlike Murnane and Phillips, Hanushek did not control for ratings of the teachers' undergraduate institutions.

Recent studies. Two additional studies appeared recently. The first study took advantage of teachers' responses to a single multiple-choice mathematics test item. Rowan, Chiang, and Miller (1997) analyzed nationally representative achievement data from the National Educational Longitudinal Study of 1988 (NELS:88). Individual students were tested in mathematics in the spring of Grades 8 and 10, and the survey given to their 10th-grade teachers included a single, high school–level mathematics test item. The researchers found that students whose teachers answered the item correctly posted larger mathematics gains between 8th and 10th grades, even after controlling for whether teachers held mathematics-related degrees.[13] This study carries much less weight, however, than it would have if teachers had completed more test items or an entire mathematics test.

A second recent study capitalized on Alabama personnel records, some of which contained teachers' ACT college entrance examination test scores. Ferguson and

Ladd (1996) used composite ACT scores. These scores combined English, mathematics, social studies reading, and natural sciences reading components (American College Testing, 1989). Their initial analysis showed that student reading score gains from 3rd grade to 4th grade were positively related to the average teacher ACT score at the students' schools. The relationship was unclear for mathematics score gains.

Ferguson and Ladd also analyzed district-level averages from 127 Alabama school districts. The authors created artificial gain scores using achievement data from 3rd and 4th graders, and 8th and 9th graders, for the 1990–1991 school year. The average differences in each district between the mathematics scores of younger students and older students were positively related to the average teacher ACT score in the district. Both the school-level analysis and the district-level analysis controlled for teacher experience and whether teachers had master's degrees.

Joint Interpretation

Examined jointly, the seven studies involving teacher test scores yielded somewhat divergent findings: Determinate findings included five positives and two negatives. Several explanations for this divergence were considered.

That these seven determinate findings could have occurred randomly—and that there is in fact no relationship between student achievement and teacher test scores—is very unlikely. And although one might instead attempt to conclude that the two negative findings were random occurrences, that explanation also seems implausible given that the negative findings come from strong studies: Both used individual student-level data, and Summers and Wolfe in particular controlled for an extensive set of schooling factors.

Another potential method of reconciling the studies, which did not point to clear conclusions, involves the content of the student and teacher tests featured in these studies. For example, with regard to the student tests, Hanushek (1992) found that teachers' performance on word tests affected their students' ITBS reading score gains but not their ITBS vocabulary score gains. This is consistent with Murnane and Phillips (1981), who reported no relationship between teachers' word test scores and student achievement gains. Future work may discover that the relationship between student gains and teacher assessment scores depends upon alignment of the underlying instruments, but the set of studies that exists today is insufficient to make such claims.

In the end, the most plausible explanation for the divergent findings emerged through examination of the controls used in each of the seven studies. Specifically, the two studies that generated negative findings both controlled for college ratings. In contrast, none of the studies yielding positive findings for teachers' test scores controlled for college ratings. The negative findings reflect the effect of test scores after controlling for the effects of college ratings, which are likely to capture a dimension of teacher quality similar to that captured by test scores. Thus, the negative findings may support the five positive findings—that students learn more from teachers with higher test scores. Test scores matter, if college ratings have not already been taken into account.

Implications

Although the finding that students learn more from teachers with higher test scores certainly militates in favor of rigorous licensure examinations for teachers, it is

important to remember that policymakers face a range of different types of licensure tests. As of 1999–2000, 34 states required teaching candidates to pass tests of basic literacy and numeracy skills. In addition, 30 states required candidates to pass tests of subject matter knowledge, and 25 states required them to pass tests of pedagogical knowledge (Youngs, Odden, & Porter, 2001). The synthesis of existing studies does not permit us to conclude which types of knowledge ought to be tested.

To provide further guidance to policymakers, there is also a need for research on the relationship between student achievement and teachers' performance on tests currently in use (as opposed to the teacher tests featured in the studies in this review, none of which remain in use). Today, many states have implemented ETS's Praxis I basic skills tests, Praxis II tests of subject matter knowledge, and/or Praxis II tests of pedagogical knowledge (all developed by ETS in the 1990s). Other states have created their own tests or contracted with companies other than ETS to develop such tests.

Finally, researchers should also take note of the recent implementation of new approaches to teacher performance assessment in Connecticut, North Carolina, Ohio, and other states. These assessments were designed to improve on existing pencil and paper tests. Student achievement studies that use such assessments would be informative to those considering these options (see, e.g., Pecheone, Rogers, & Moirs, 2001).

These various recommendations resonate with those offered in the final report of a panel recently convened by the National Research Council to examine the issues of teacher licensure examinations and teacher quality. The panel's 2001 report called for "a multi-discipline, multiple methods research program" that would "examine licensure testing, beginning teacher performance, and student learning" (National Research Council Committee on Assessment and Teacher Quality, 2001, p. 29). Hopefully, this review can inform the initiation of such a research program.

Degrees and Coursework

Relevant Studies and Findings

Until recently, lack of data prevented researchers from determining whether students learned more from teachers with particular degrees or coursework. The available data sets contained information on teachers' degree level (e.g., bachelor's, master's, etc.), and results were mixed. Most studies were indeterminate (Harnisch, 1987; Hanushek, 1992; Link & Ratledge, 1979; Murnane, 1975; Murnane & Phillips, 1981; Rivkin, Hanushek, & Kain, 2001; Summers & Wolfe, 1975a, 1977), and the four determinate findings were both positive (Ferguson & Ladd, 1996) and negative (Eberts & Stone, 1984; Ehrenberg & Brewer, 1994; Kiesling, 1984). Close examination of the four determinate studies does not permit reconciliation of their findings. Although three suggest a negative influence on student learning, the positive results reported by Ferguson and Ladd (1996) are convincing. Moreover, as is clear from the information presented in the Appendix, neither differences in controls nor differences in the specification of degree level could explain the divergence. As a result, no conclusions were possible.

The recent improvement in data collection on degrees and coursework led to results that make apparent that the earlier, mixed results for degree level were at least partly attributable to the failure of those studies to identify whether the additional degree was related to the subject being taught. This finding has been documented

most clearly by researchers using NELS:88, which was briefly discussed in the preceding section on teachers' test scores. NELS:88 is a nationally representative survey of about 24,000 Grade 8 students conducted in the spring of 1988. A subset of these students was surveyed again in the spring of 10th (1990) and 12th (1992) grades. At the time of each survey, students took one or more subject-based tests in mathematics, science, English/writing, and history. Therefore, the NELS:88 follow-up data sets permit longitudinal analyses of growth in student achievement from 8th to 10th grade, 10th to 12th grade, and 8th to 12th grade in particular subjects. The NELS:88 data also include information on relevant student, teacher, and school characteristics.

Three analyses take advantage of the detailed teacher data in NELS:88 on degrees. The analysis by Goldhaber and Brewer (1997a) illustrates the key finding most clearly. No differences were evident when the authors examined whether 10th-grade mathematics students scored better when their teachers had master's degrees. However, introducing information about the subject of the teachers' degrees produced significant results. Mathematics students whose teachers had master's degrees in mathematics had higher achievement gains than those whose teachers had either no advanced degrees or advanced degrees in nonmathematics subjects. In addition, students whose teachers had bachelor's degrees in mathematics learned more than students whose teachers had bachelor's degrees in nonmathematics subjects. The contributions of these two indicators of subject preparation were independent of several other teacher characteristics; the student achievement model controlled for certification, mathematics certification, and years of high school teaching experience.

The two other analyses of the NELS:88 data replicated this finding, again for mathematics, but with slight variations. Goldhaber and Brewer (2000) used the 12th- rather than 10th-grade students and again found that students learned more from teachers with mathematics majors and from teachers with master's degrees in mathematics. Rowan, Chiang, and Miller (1997) used a single variable to indicate whether the teacher had an undergraduate or graduate degree in mathematics. In addition, as described earlier, they added a crude control for teachers' test scores in mathematics knowledge.

If having a degree in mathematics makes a teacher more effective, one might expect that measures of coursework in mathematics would also predict effectiveness. Two studies used such measures. Eberts and Stone (1984) recorded the number of college-level, mathematics-related courses taken by teachers in the previous 3 years. No relationship to 4th graders' mathematics achievement was apparent.

The second study employing measures of teachers' mathematics course taking did conclude that some relationships existed. Monk and King (1994) used achievement data from the Longitudinal Survey of American Youth (LSAY), a study that followed a national probability sample of 2,831 public school 10th graders from fall 1987 into their senior year. Students took math and science tests each fall from 1987 to 1989 that were based on the National Assessment of Educational Progress (NAEP).

The LSAY identified the science and mathematics teachers of each student, and a survey captured teachers' experience and course taking. To develop the course taking measure, the authors pooled all undergraduate and graduate courses into three categories: mathematics, life science, and physical science.

Although the study by Monk and King yielded many indeterminate findings, they noted a handful of positive relationships between mathematics gains and course

taking in mathematics. First, controlling for teacher experience, the 10th-grade students who performed well in the fall test posted higher 1-year gains when their mathematics teachers had more mathematics courses. Second, observing the gain between 1987 and 1989, the authors found that students learned more mathematics when their 10th-grade and 11th-grade mathematics teachers had taken more mathematics courses.

The results of another study based on data from the LSAY (Monk, 1994) are often cited in discussions of student achievement and teacher characteristics. However, these results were excluded from this review owing to the absence of controls for students' socioeconomic status.[14]

In mathematics, then, degrees and coursework appear related to teacher effectiveness, but what about other subjects? In other subjects, student achievement results have been indeterminate or inconsistent. Goldhaber and Brewer (1997a) performed additional analyses for science, English, and history. The authors discerned positive effects on 10th graders' achievement gains for science teachers' bachelor's degrees, but they did not report any other relationships. In their later analysis of 12th-grade students' gains (Goldhaber & Brewer, 2000), the authors again found positive effects in science, but this time the results were not statistically significant. Monk and King (1994) also examined science achievement gains and generated only one determinate result: The juniors in their sample learned *less* from teachers with more physical science coursework.

Joint Interpretation

To join these various findings, it is reasonable to treat each subject separately. In history and English, with no determinate findings, this review cannot draw any conclusions about the importance of degrees and coursework. In science, we confront only two determinate findings, and they point in opposite directions. Science includes both physical and life science, so one plausible explanation for the contradiction is that measurement of degrees, course taking, and teaching within science is not yet sufficiently specific. Research with more fine-grained data collection strategies is necessary when science achievement is studied.

In mathematics, all determinate findings were positive, so it is possible to assert that students learn more from teachers with more mathematics-related coursework and degrees. However, all of the positive determinate findings focus on high school students. The indeterminate finding by Eberts and Stone (1984) involved elementary school students. Therefore, this review can conclude that high school students learn more mathematics when their mathematics teachers have additional degrees or coursework in mathematics. More evidence would be needed to make that assertion for elementary school students.

Implications

Almost without exception, U.S. school districts' teacher compensation systems reward teachers for holding advanced degrees. The student achievement studies reviewed above do not refute the possibility that such a policy is wise, but it is certainly clear in the case of mathematics that alignment between degree content and subject assignment is important.

Another relevant leverage point for policymakers is their power to specify coursework and degree requirements for different assignments. Many states have such

requirements. The research reviewed here indicates that increased requirements accompanied by increased compensation would likely have payoffs in terms of student achievement, at least in the case of mathematics. Whether these payoffs occur will ultimately depend on the costs to prospective teachers of completing these requirements as well as the relationship's effect size, which remains unclear for the reasons given above in the Methods section.

Finally, it is interesting to note that these studies leave some uncertainty with respect to whether students learn more from teachers with subject-related education degrees (e.g., mathematics education, as opposed to mathematics). None of the NELS:88 teacher surveys included such response categories. Those with subject-related education degrees probably selected "education," but it is likely that at least some would have chosen the subject to which their degree was related. Monk and King (1994) did not report how they handled the distinction.[15] Eberts and Stone (1984) labeled their course counts as courses "related to teaching mathematics" and therefore probably included both mathematics courses and mathematics education courses. In sum, studies have not been sufficiently aggressive about distinguishing the two. For researchers, these findings leave a clear roadmap for future research. Studies that use subject-specific measures of teacher preparation and that distinguish between subject preparation and preparation in the methods of teaching a subject would provide valuable new information.

Certification Status

Relevant Studies and Findings

As was true for teacher degrees, the effects of teacher certification appear only when teachers have certification *for the subject taught,* and these findings have been in mathematics. Only two studies meeting this review's design standards—both by Goldhaber and Brewer (1997, 2000)—have examined certification; other studies reported in the literature either are not longitudinal or do not control for students' socioeconomic status.[16]

In their first study, described in more detail above, Goldhaber and Brewer (1997a) tested two certification indicators. The first simply asked whether the teacher was certified, without reference to any particular subject, and yielded only one determinate relationship: Students taking English classes appeared to learn *less* from English teachers who held certification.

The second model used by Goldhaber and Brewer (1997a) added information about the particular subject in which teachers claimed certification. The results for English became indeterminate, suggesting that the earlier negative finding was caused by English teachers holding certification outside of English. Results for history were also indeterminate. However, the authors reported that mathematics students had higher achievement gains when their teachers held certification in mathematics as compared with holding no mathematics certification—which includes teachers who hold no certification at all and teachers who hold certification to teach in other areas. The same comparison applied to science achievement gains also revealed a positive relationship.

Notably, this study did not report how differences in certification type were treated. For a given subject, most states offer a variety of types of certification, and there is uncertainty among researchers and policymakers about the comparative effectiveness of teachers whose certification is of a nonstandard type (e.g., emer-

gency, provisional, temporary). Proponents of "alternative certification" argue that some trimming down of the standards attracts persons with greater qualities by reducing entry barriers.

In a second study that employed subject-specific teacher certification variables, Goldhaber and Brewer explicitly addressed differences in certification type. As described earlier, Goldhaber and Brewer (2000) focused on the gains made by the NELS:88 students between the 10th and 12th grades. The survey administered to the teachers of 12th-grade students included an item that asked about certification type in mathematics and certification type in science. The authors therefore examined certification type in connection with mathematics and science achievement gains only. The determinate finding from their analysis was that students' mathematics gains were higher when their mathematics teachers held standard certification in mathematics, as compared with the gains of those whose teachers held either (a) no certification in mathematics (which includes teachers certified in other fields as well as teachers with no certification in any subject) or (b) private school certification in mathematics.

This study also included further analysis of certification type, which spawned additional dialogue (see Darling-Hammond et al., 2001; Goldhaber & Brewer, 2001). Goldhaber and Brewer (2000) compared the mathematics gains of students whose teachers held standard certification in mathematics with those who had checked a certification type labeled as follows: "temporary, provisional, or emergency certification (requires additional coursework before regular certification can be obtained)." Controlling for the other variables in their model, the authors were not able to discern differences and interpreted this indeterminate finding as a suggestion that the teachers in this comparison were equally effective. However, the conservative standard adopted by this review precludes drawing a firm conclusion on the basis of an indeterminate finding.[17]

Finally, another study some analysts will characterize as relevant to the debate about certification type was conducted by Raymond, Fletcher, and Luque (2001). This study used data from the Houston Independent School District to address whether students learn more from teachers who secured their jobs through Teach for America (TFA). However, the comparison made in their study was *not* one between TFA teachers and teachers with standard certification for their subjects. Instead, the study compared the effectiveness of TFA teachers against *all* other beginning teachers, controlling for other factors. Those beginning teachers included some with standard certification in their subjects, some with nonstandard certification types, and some with no certification whatsoever. The study therefore informs the decision by the Houston school district regarding whether or not to accept TFA teachers but does not allow generalizations about certification type.

Joint Interpretation

The study descriptions presented in the Appendix show that these two studies have reasonably strong design features. Both are individual-level analyses, and both use a substantial set of socioeconomic status controls. Controls for other school inputs and other teacher variables were less extensive than in some other studies but nevertheless present. Because the first study left unclear how certification type was treated and the second study was able to draw conclusions about standard certification only, we conclude that mathematics students learn more when their teachers

have standard mathematics certification (as compared with private school mathematics certification or no mathematics certification).

Implications

The findings for certification mirror those for degrees and coursework. In short, subject-specific measures matter. The finding that math teachers with standard mathematics certification outperform those with no mathematics-related certification indicates that states' teaching standards in mathematics are, on average, meaningful.

In drawing policy inferences from this finding, one must not lose sight of the fact that each state has its own requirements that must be met in order to achieve standard certification in mathematics (e.g., subject-related coursework, passage of licensure tests, etc.). The finding does not point us to the specific requirements that are important. Whether streamlining these requirements would result in changes in quality is not answered by existing research.

To produce findings with greater clarity in regard to certification, researchers need to design studies that take into consideration the particular requirements associated with particular certification types used in individual states. The Schools and Staffing Survey, a federally funded study that describes patterns of certification but does not measure student achievement, recently began to elicit somewhat more detailed information about teacher preparation.[18] For the purposes of studying student achievement and certification, one solution would be to focus on particular states or small numbers of states that share common definitions and requirements for terms such as probationary certification, emergency certification, and so forth.

In addition, because the goal of reforms to certification systems would be to increase the average quality of the teachers who supply themselves to the profession, parallel research efforts are needed to examine differences in retention rates among teachers with different types of certification. These recommendations are consistent with those offered by others who have produced related analyses.[19]

Other Characteristics

The characteristics addressed in the sections above on college ratings, test scores, degrees and coursework, and certification are those characteristics for which research has proven sufficiently conclusive to inform policymakers. In the case of many other characteristics, research either does not exist or has not resulted in clear findings. This section identifies those characteristics and discusses each briefly.

Perhaps the most conspicuous characteristic absent from this review is teacher experience. Of the 21 qualifying student achievement studies, 19 used information about how many years teachers had been teaching, and most of the determinate results were positive. However, for reasons first identified by Murnane and Phillips (1981, pp. 94–97), we decided that findings regarding experience were too difficult to interpret. First, experience necessarily captures the effect of whether teachers were hired during a shortage or a surplus. Therefore, meaningful generalization would require controls for shortage and surplus conditions for each possible year of hire.

Second, experience measures capture differences in teacher motivation resulting from time constraints on parents during years when child rearing requires more attention. Meaningful generalization would require knowing whether the teacher had dependent children at the time.

Finally, if there are differences in effectiveness between those who leave the profession and those who stay, experience measures would capture those as well, and such differences are probably dynamic—changing with cultural trends as well as labor market conditions. Thus, while the myriad effects captured by experience make it a valuable control variable, the relationships that emerge between experience and student achievement are difficult to interpret.

Another teacher characteristic not discussed in this review was teacher race. Race was excluded from this review in order to maintain focus on clear findings. The findings on teacher race—and specifically on whether students learn more from teachers of their own race or ethnicity—have been mixed (see Ferguson, 1998). If clear findings are ever achieved, drawing implications from those findings will require careful analysis.

This review also did not discuss the importance of holding degrees in education. Studies have not explicitly distinguished between degrees in subjects and degrees in the teaching of particular subjects, nor have they distinguished between degrees in the teaching of particular subjects and general degrees in teaching or education. Studies that make these distinctions would fill an important gap in the literature.

While student achievement studies have not yet assessed the importance of the characteristics identified in this section, theory and intuition suggest that they may be important. Therefore, studies that find relationships with other teacher characteristics and use these characteristics as controls would be more conclusive than those that do not.

Conclusion

We approached the task of review in a way that would create a clear interpretation of the research for policymakers and researchers interested in the relationship between teacher characteristics and student achievement gains. Our methods were systematic and transparent. More specifically, (a) our scope focused us on compelling findings; (b) we considered each relevant study's unique strengths and weaknesses; and (c) we reconciled findings and joined them together through deliberate reasoning.

The interpretations rendered by this review are easily summarized. The studies confirm that students learn more from teachers with certain characteristics. In the case of teachers' college ratings and test scores, positive relationships exist and should be investigated further to learn about the relative importance of specific college characteristics and tested skills and knowledge. In the case of degrees, coursework, and certification, findings have been inconclusive except in mathematics, where high school students clearly learn more from teachers with certification in mathematics, degrees related to mathematics, and coursework related to mathematics.

Although this added confirmation is meaningful in policy debates, there are many important questions that remain unanswered. It is possible, for instance, that results would differ if outcomes such as graduation rates or future earnings were examined. In addition, for many potentially salient teacher characteristics—such as experience, race, and study of teaching methods—studies that use convincing research designs simply do not exist or have not been conclusive. Furthermore, it is also important to consider unobservable changes in the composition of the teaching force over time. The makeup of the teaching force is certainly influenced by many factors and has changed considerably during the past three decades.[20]

Finally, for those characteristics found to be related to student achievement gains, existing studies do not offer convincing indications of effect sizes. What if those effect sizes are extremely small? Much remains open to policymakers' intuition.

Effect size estimates are in fact central to the current debate in teacher policy. An important school of thought in policy-making on teacher quality holds that using teacher characteristics in policy design is a bad idea because most of the variation in teacher quality is unseen. In other words, teachers differ greatly in their effectiveness, but teachers with and without different qualifications differ only a little. Therefore, according to this school of thought, policies that emphasize motivating principals and increasing their discretion over hiring should replace policies that require particular qualifications.

The only study that has explicitly addressed itself to generating accurate estimates of effect size is Rivkin, Hanushek, and Kain (2001). Using student achievement at *more than* two points in time, the authors attempt to remove the influences of particular students and schools that do not vary over time. This method promises to better isolate the effects of teachers but should be considered developmental until the research community has an opportunity to evaluate it. Another promising approach to generating better estimates of effect size is research involving random assignment. For instance, a new effort under way to study TFA utilizes random assignment of students to TFA and non-TFA teachers (Decker & Mayer, 2002). Such designs are not without their own difficulties (see Ehrenberg et al., 2001). They are, however, more transparent than quasi-experimental studies, and the perspective yielded on effect sizes would do much to advance debates on teacher policy.

Notes

Some of the findings presented in this article appear in an introductory chapter of Wayne (2000), a doctoral dissertation. Although the responsibility for errors belongs solely to the authors, we would like to thank Dale Ballou, Linda Darling-Hammond, William Galston, Daniel Goldhaber, Willis Hawley, Mark Lopez, and two anonymous reviewers for helpful comments on earlier versions.

[1] The empirical research that supports this claim is a body of student achievement studies that ignore the particular characteristics of teachers and focus instead on variations in student achievement gains from one teacher to another. See Goldhaber and Brewer (1997b); Jordan, Mendro, and Weerasinghe (1997); Hanushek (1971, 1992); Murnane (1975); Murnane and Phillips (1981); Rivkin, Hanushek, and Kain (2001); Sanders and Rivers (1996); and Wright, Horn, and Sanders (1997).

[2] Differences in interpretation are evident in several recent treatments, including Ballou and Podgursky (1999, 2000); Darling-Hammond (1999b, 2001); Darling-Hammond, Berry, and Thoreson (2001); Goldhaber and Brewer (2001); and Walsh (2001).

[3] Certification requirements vary considerably across states. In most states, candidates for teaching must earn a minimum grade point average and/or achieve a minimum score on tests of basic skills, general academic ability, or general knowledge in order to be admitted to teacher education or gain a credential. In addition, candidates in many states must complete a major or minor in the subject(s) to be taught and/or pass a subject matter test, take specific courses in education, and/or pass a test of teaching knowledge and skill. In 2001–2002, 37 states required candidates to pass tests of basic skills or general knowledge, 33 states required candidates to pass tests of content knowledge, and 26 required candidates to pass tests of pedagogy.

[4] The payoff to trying to identify relevant conference papers and dissertations is quite small relative to the effort required, especially given the time required to obtain such documents, and one would expect that good work would eventually appear as regular publications.

[5] Readers interested in learning more about the studies that were excluded should consult the appendices available in Greenwald, Hedges, and Laine (1996) as well as Mitchell et al. (2001).

[6] The study is Grissmer et al. (2000), which analyzed trends in states' average NAEP scores. Another study of state NAEP averages by Darling-Hammond (1999a) had already been excluded because it examined averages at a point in time rather than longitudinally.

[7] Early work on this section was completed by Andrew J. Wayne while working at the National Partnership for Excellence and Accountability in Teaching, which was funded through the U.S. Department of Education, Office of Educational Research and Improvement, under Contract RD97124001.

[8] Several researchers have noted the discriminatory impact of teacher tests on minority candidates, particularly in southern states (Goertz & Pitcher, 1985; Graham, 1987; Smith, Miller, & Joy, 1988; Ludlow, 2001). The use of tests developed by National Evaluation Systems for Alabama was prohibited in the 1980s owing to psychometric reasons related to their discriminatory impact on African American candidates (Ludlow, 2001).

[9] Strauss and Sawyer (1986) also examined the relationship between student achievement and NTE scores. That study was excluded, however, because it did not use pretest scores.

[10] There is a potential for confusion and, perhaps, debate on this point as a result of a third publication by the authors, Summers and Wolfe (1975b). That publication notes two positive findings not noted in the other publications: 8th graders learned more from social studies teachers with higher NTE social studies scores, and 12th graders with above average achievement scores learned more from English teachers with higher NTE English scores. Still, we believe it is appropriate to summarize these results as simply "indeterminate," which we note is consistent with the treatment in Summers and Wolfe (1975a), a more academically oriented analysis of the results for NTE scores across all three grade levels. That publication does not discuss the relationship of NTE area scores among its significant findings. Of course, whether ours is the best possible summary of the results for NTE area scores is of minor importance, given the interpretation reached later in this article regarding test scores.

[11] To clarify the areas of knowledge and skills measured by the TECAT, we contacted officials at the Texas Education Agency and the Texas State Board for Educator Certification and examined the registration bulletin and study guide for the TECAT. We concluded that while this test evaluates reading and writing skills, it clearly also integrates professional knowledge into these items. In addition, we contacted Ron Ferguson in 1998, and we learned that his analyses used TECAT reading scores specifically. Therefore, we interpret the Ferguson results as applicable to a test that measured teachers' reading skills and professional knowledge.

[12] Ferguson analyzed data from almost 900 school districts in Texas, out of more than 1,000 districts in the state. Most of the districts omitted from the study owing to missing data were very small. In addition, Dallas and Houston were not included in the analysis because the weighting scheme in the estimate procedure would have given these two districts too much influence over the results.

[13] Readers will note that this study models a student's 2-year gain as a function of the quality of a teacher who teaches the student for less than 1 year. Although a noticeably incomplete model, it is nevertheless sufficient. The exposure to teachers of unknown quality introduces measurement error, which in regression models would make it *more* difficult to discern a statistically significant relationship between the independent and dependent variables. A relationship appears in this study and others that use the same data set *despite* the added measurement error.

[14] The Monk (1994) study contains no mention of any controls for socioeconomic status, and in a personal communication Monk confirmed that no such controls were used. Had it been admissible, it would have helped to confirm the conclusion reached in this section that mathematics-related preparation improves the effectiveness of mathematics teachers.

[15] The Monk (1994) study, which was excluded as a result of the absence of controls for socioeconomic status, explicitly separated education-related courses in a subject from other courses in a subject.

[16] Readers may be familiar with a study by Hawk, Coble, and Swanson (1985) that concluded that students learn more from certified teachers. That study did not control for student socioeconomic status and was therefore excluded from this review. The study by Fetler (1999) is often cited as well, but it was excluded owing to the fact that it did not use pretest scores.

[17] As a result of space limitations, we are not able to summarize all areas of agreement and disagreement between Goldhaber and Brewer (2000) and Darling-Hammond et al. (2001), but it is important to note that they are in agreement that the teachers in the study with nonstandard certification otherwise had qualifications that were on average quite similar to those of the teachers with standard certification. This fact raises important questions about the measure used to distinguish teachers and thus gives us reason to await further research before drawing conclusions about certification type.

[18] Documentation on the Schools and Staffing Survey, including the instruments given to teachers, is available online at *www.nces.ed.gov/surveys/sass.*

[19] We are in agreement with Goldhaber and Brewer (2001) and Darling-Hammond et al. (2001) that research on certification status and student achievement is complicated because (a) certification is defined differently in different states and (b) definitions of certification within a single state often change over time. Further, both groups of researchers provide useful recommendations for future research on teacher certification policy. In particular, Goldhaber and Brewer call for "more detailed information on the specific date, location, and content of an individual teacher's certification experience and more comprehensive information on how state licensing policy has changed over time" (2000, p. 141). For their part, Darling-Hammond et al. recommend a research agenda that includes efforts "to understand how different teacher certification strategies encourage or discourage the construction of programs that produce well prepared teachers who stay in the profession, and how state policies distribute well prepared teachers equitably to all children in the state, regardless of race and income" (2001, p. 71).

[20] Since the 1970s, women and racial minorities have had many more professional opportunities outside of education. There were relatively few teaching positions available in the 1970s, owing to student and teacher demographics, which also affected college students' attitudes about teaching. See generally Warren (1989) and Sedlak and Schlossman (1986). In addition, there is an ongoing debate about the academic qualifications of prospective teachers. See Gitomer, Latham, and Ziomek (1999) and Henke, Chen, and Geis (2000).

References

American College Testing. (1989). *The enhanced ACT assessment: Using concordance tables.* Iowa City, IA: Author.

Ballou, D., & Podgursky, M. (1999, October 13). Reforming teacher preparation and licensing: What is the evidence? *Teachers College Record.* Retrieved December 7, 1999, from *http://www.tcrecord.org*

Ballou, D., & Podgursky, M. (2000, May). Reforming teacher preparation and licensing: Continuing the debate. *Teachers College Record.* Retrieved June 15, 2000, from *http://www.tcrecord.org*

Coleman, J. S. (1966). *Equality of educational opportunity.* Washington, DC: Office of Education, U.S. Department of Health, Education, and Welfare.

Darling-Hammond, L. (1999a). *Teacher quality and student achievement: A review of state policy evidence.* Seattle: Center for the Study of Teaching and Policy, University of Washington.

Darling-Hammond, L. (1999b, October 13). Reforming teacher preparation and licensing: Debating the evidence. *Teachers College Record.* Retrieved December 7, 1999, from *http://www.tcrecord.org*

Darling-Hammond, L. (2001). *The research and rhetoric on teacher certification: A response to "Teacher certification reconsidered."* New York: National Commission on Teaching and America's Future.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23,* 57–77.

Decker, P., & Mayer, D. (2002). *Studying alternative certification using random assignment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Dewey, J., Husted, T. A., & Kenny, L. W. (2000). The ineffectiveness of school inputs: A product of misspecification? *Economics of Education Review, 19,* 27–45.

Eberts, R. W., & Stone, J. A. (1984). *Unions and public schools: The effect of collective bargaining on American education.* Lexington, MA: Heath.

Ehrenberg, R. G., & Brewer, D. J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond. *Economics of Education Review, 13,* 1–17.

Ehrenberg, R. G., & Brewer, D. J. (1995). Did teachers' verbal ability and race matter in the 1960s? *Coleman* revisited. *Economics of Education Review, 14,* 1–21.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Wilms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest, 2*(1), 1–30.

Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *Industrial and Labor Relations Review, 48,* 547–561.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28,* 465–498.

Ferguson, R. F. (1998). Can schools narrow the Black-White test score gap? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 318–374). Washington, DC: Brookings Institution.

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. 265–298). Washington, DC: Brookings Institution.

Fetler, M. (1999). High school staff characteristics and mathematics test results. *Education Policy Analysis Archives, 7*(9). Retrieved March 3, 2000, from http://epaa.asu.edu/epaa/v7n9.html

Gitomer, D. H., Latham, A. S., & Ziomek, R. (1999). *The academic quality of prospective teachers: The impact of admissions and licensure testing.* Princeton, NJ: Educational Testing Service.

Goertz, M. E., & Pitcher, B. (1985). *The impact of NTE use by states on teacher selection.* Princeton, NJ: Educational Testing Service.

Goldhaber, D. D., & Brewer, D. J. (1997a). Evaluating the effect of teacher degree level on educational performance. In W. J. Fowler (Ed.), *Developments in school finance, 1996* (pp. 197–210). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Goldhaber, D. D., & Brewer, D. J. (1997b). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, 32,* 505–523.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school certification status and student achievement. *Educational Evaluation and Policy Analysis, 22,* 129–146.

Goldhaber, D. D., & Brewer, D. J. (2001). Evaluating the evidence on teacher certification: A rejoinder. *Educational Evaluation and Policy Analysis, 23,* 79–86.

Graham, P. A. (1987). Black teachers: A drastically scarce resource. *Phi Delta Kappan, 68,* 598–605.

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

Grissmer, D. W., & Flanagan, A. (2000). Moving educational research toward scientific consensus. In D. W. Grissmer & J. M. Ross (Eds.), *Analytic issues in the assessment of student achievement* (pp. 43–90). Washington, DC: U.S. Department of Education.

Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us.* Santa Monica: RAND.

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 14). Washington, DC: American Educational Research Association.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro-data. *American Economic Review, 61,* 280–288.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100,* 85–117.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19,* 141–164.

Hanushek, E. A., Gomes-Neto, J. B., & Harbison, R. W. (1996). Efficiency-enhancing investments in school quality. In N. Birdsall & R. H. Sabot (Eds.), *Opportunity foregone: Education in Brazil* (pp. 385–424). Washington, DC: Inter-American Development Bank/Johns Hopkins University Press.

Harnisch, D. L. (1987). Characteristics associated with effective public high schools. *Journal of Educational Research, 80,* 233–241.

Hawk, P., Coble, C. R., & Swanson, M. (1985). Certification: It does matter. *Journal of Teacher Education, 36*(3), 13–15.

Henke, R. R., Chen, X., & Geis, S. (2000). *Progress through the teacher pipeline: 1992–93 college graduates and elementary/secondary school teaching as of 1997.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Hussar, W. J. (1999). *Predicting the need for newly hired teachers in the United States to 2008–09.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Ingersoll, R. M. (1996). *Out-of-field teaching and educational equity.* Washington, DC: U.S. Department of Education.

Jordan, H. R., Mendro, R., & Weerasinghe, D. (1997). *Teacher effects on longitudinal student achievement.* Paper presented at the annual meeting of the Center for Research on Educational Accountability and Teacher Education, Indianapolis, IN.

Kain, J. F., & Singleton, K. (1996, May–June). Equality of educational opportunity revisited. *New England Economic Review,* pp. 87–111.

Kiesling, H. J. (1984). Assignment practices and the relationship of instructional time to the reading performance of elementary school children. *Economics of Education Review, 3,* 341–350.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research.* Princeton, NJ: Princeton University.

Link, C. R., & Mulligan, J. G. (1986). The merits of a longer school day. *Economics of Education Review, 5,* 373–381.

Link, C. R., & Ratledge, E. C. (1979). Student perceptions, I.Q. and achievement. *Journal of Human Resources, 14,* 98–111.

Ludlow, L. (2001). Teacher test accountability: From Alabama to Massachusetts. *Education Policy Analysis Archives, 9*(6). Retrieved April 6, 2002, from http://epaa.asu.edu/epaa/v9n6.html

Ludwig, J., & Bassi, L. J. (1999). The puzzling case of school resources and student achievement. *Educational Evaluation and Policy Analysis, 21,* 385–403.

Maynard, R., & Crawford, D. (1976). School performance. In D. L. Bawden & W. S. Harrar (Eds.), *Rural income maintenance experiment: Final report* (Vol. VI, Part II, pp. 1–104). Madison: University of Wisconsin, Institute for Research on Poverty.

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality.* Washington, DC: National Academy Press.

Monk, D. H. (1994). Subject area preparation of secondary math and science teachers and student achievement. *Economics of Education Review, 13,* 125–145.

Monk, D. H., & King, J. (1994). Multilevel teacher resource effects on pupil performance in secondary mathematics and science: The case of teacher subject-matter preparation. In R. Ehrenberg (Ed.), *Contemporary policy issues: Choices and consequences in education* (pp. 29–58). Ithaca, NY: ILR.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger.

Murnane, R. J., & Phillips, B. R. (1981). What do effective teachers of inner-city children have in common? *Social Science Research, 10,* 83–100.

Pecheone, R., Rogers, J., & Moirs, K. (2001). *A beginning teacher survey study: A policy-practice perspective.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Raymond, M., Fletcher, S. H., & Luque, J. (2001). *Teach for America: An evaluation of teacher differences and student outcomes in Houston, Texas.* Stanford, CA: Center for Research on Education Outcomes.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2001). *Teachers, schools, and academic achievement.* Amherst, MA: Amherst College.

Rowan, B. (2002). *What large scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools.* Ann Arbor: University of Michigan.

Rowan, B., Chiang, F.-S., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education, 70,* 256–284.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future academic achievement.* Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sedlak, M., & Schlossman, S. (1986). *Who will teach? Historical perspectives on the changing appeal of teaching as a profession.* Santa Monica, CA: Rand.

Smith, G. P., Miller, M. C., & Joy, J. (1988). A case study of the impact of performance-based testing on the supply of minority teachers. *Journal of Teacher Education, 39*(4), 45–53.

Strauss, R. P., & Sawyer, E. A. (1986). Some new evidence on teacher and student competencies. *Economics of Education Review, 5,* 41–48.

Summers, A. A., & Wolfe, B. L. (1975a). *Equality of educational opportunity quantified: A production function approach.* Philadelphia: Federal Reserve Bank of Philadelphia, Department of Research.

Summers, A. A., & Wolfe, B. L. (1975b). *Which school resources help learning? Efficiency and equity in Philadelphia public schools.* Philadelphia: Federal Reserve Bank of Philadelphia.

Summers, A. A., & Wolfe, B. L. (1977). Do schools make a difference? *American Economic Review, 67,* 639–652.

Turner, S. E. (2000). A comment on "Poor school funding, child poverty, and mathematics achievement." *Educational Researcher, 29*(5), 15–18.

U.S. Department of Education. (2002). *Meeting the highly qualified teachers challenge.* Washington, DC: Author.

Walsh, K. (2001). *Teacher certification reconsidered: Stumbling for quality.* Baltimore: Abell Foundation.

Warren, D. (1989). *American teachers: Histories of a profession at work.* New York: Macmillan.

Wayne, A. J. (2000a). *Federal policies to improve teacher quality for low-income students.* Unpublished doctoral dissertation, University of Maryland, College Park.

Wayne, A. (2000b). Teacher supply and demand: Surprises from primary research. *Education Policy Analysis Archives, 8*(47). Retrieved September 19, 2000, from http://epaa.asu.edu/epaa/v8n47/

Wayne, A. (2002). Teacher inequality: New evidence on disparities in teachers' academic skills. *Education Policy Analysis Archives, 10*(30). Retrieved June 30, 2002, from http://epaa.asu.edu/epaa/v10n30/

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations.* Seattle, WA: Center for the Study of Teaching and Policy.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11,* 57–67.

Youngs, P., Odden, A., & Porter, A. C. (2001). *State leadership in teacher licensure.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Youngs, P., Odden, A., & Porter, A. C. (in press). State policy related to teacher licensure. *Educational Policy.*

Authors

ANDREW J. WAYNE is an Education Researcher at SRI International, 1100 Wilson Blvd., Suite 2800, Arlington, VA 22209-3915; e-mail *andrew.wayne@sri.com.* His areas of specialization are teacher quality and educational technology.

PETER YOUNGS is a Research Associate at Stanford University, School of Education, CERAS Building, Room 109-O, Stanford, CA 94305; e-mail *pyoungs@stanford.edu.* His area of specialization is state policy related to teacher education, licensure, and induction.

APPENDIX

Aggregation level and description of achievement data, teacher quality variables, and other controls for each qualifying study

Study citation	Achievement data	Teacher quality variables	Other controls	Additional features
Eberts and Stone (1984)	*Student:* Test scores in mathematics from the Sustaining Effects Survey for 14,959 elementary school students	*Teacher:* Years of teaching experience; highest degree earned (less than a BA, BA, MA or hours equivalent, PhD), number of college-level courses related to teaching math taken in the last 3 years, hours of formal on-the-job mathematics training ("in-service" training) in the last 3 years	*Student:* Gender, race, early childhood education (index), parents' participation in school activities (index), economic status (Orshandsky Poverty Index) *Teacher:* Time in instruction, time in preparation, time in administrative duties *School:* Ratios of administrators to students, teacher to students, clerical staff to students, principal highest degree, principal experience, principal teaching experience, principal hours per year doing leadership activities	
Ehrenberg and Brewer (1994)	*Student:* Composite test scores (math, reading, vocabulary) from HS&B for 2,650 secondary students, including 1,543 White students, 254 Black students, and 463 Hispanic students	*School:* Percentage of teachers with 10+ years of experience, master's degree; average selectivity of teachers' undergraduate institutions (6 categories from Barron's)	*Student:* Sex, single parent, parent's education, family income, family size *School:* School averages for student sex, single parent, parent's education, family income, family size; school urbanicity; students percentage Black, students percentage Hispanic, students percentage lowest SES quartile, faculty percentage Black, faculty percentage Hispanic, teacher-student ratio *District:* Expenditures per pupil	IV Corrected for dropouts

(*continued*)

APPENDIX (*Continued*)

Study citation	Achievement data	Teacher quality variables	Other controls	Additional features
Ehrenberg and Brewer (1995)	*School:* Average composite scores (verbal aptitude, nonverbal aptitude, reading, and math) from the EEOC study for 969 elementary schools and 256 high schools	*School:* Percentage with master's degree, average score on test of verbal aptitude, average years of experience	*School:* Students percentage female, students percentage Black, students percentage no father or no mother in the home, students percentage have a telephone in the home, students percentage receive free lunches, mean income of the students' families, mean education levels of the fathers and mothers, urbanicity, number of books per pupil in the school's library, pupil-teacher ratio of the school, percentage of teachers Black and White	IV
Ehrenberg, Goldhaber, and Brewer (1995)	*Student:* Tenth-grade students' test scores from the NELS:88: 5,113 in mathematics, 4,357 in science, 6,196 in English, 2,943 in history	*Teacher:* Years of experience, certified in subject, unspecified subject matter background variables, unspecified degree-level variables *School:* Percentage of teachers with at least a master's degree	*Student:* Sex, race/ethnicity, parent education, family size, family income, whether student was learning disabled or had limited English proficiency *Teacher/classroom:* Class size, class percentage minority, teacher race/ethnicity, teacher gender *School:* Total enrollment, percentage of graduates who enroll in college, racial distribution of student body, percentage of teachers with at least a master's degree, highest salary paid to teachers	
Ferguson (1991) and Ferguson (1998)	*District:* TEAMS scores in reading and math for 890 school districts, in odd grade levels (1, 3, . . . 11)	*District:* Percentage 5–9 years of experience, percentage 9+ years of experience, percentage master's degrees, average TECAT score	*District:* Students per teacher, students per primary school, students per high school, students per district, adults' education levels, students percentage in poverty, students percentage from female-headed households, students percentage English	

			second language, students percentage Hispanic, students percentage Black, students percentage in public school, whether borders poor districts, urbanicity, salary differential in nearby districts
Ferguson and Ladd (1996): district level	*District:* Stanford Achievement Test (SAT) scores for eighth graders and Basic Competency Test (BCT) math scores for ninth graders in 127 districts	*District:* College entrance ACT scores, percentage with 5 years experience, percentage with master's degrees	*District:* Class size, parental education, per capita income, percentage of families with school-age children below the poverty line, adults' education levels, per capita income (natural logarithm), poverty rate for families with school-age children, percentage students non-White, district enrollment, public school students as a percentage of all students, percentage of district that is urban, whether district is a city
Ferguson and Ladd (1996): student level	*Student:* Stanford Achievement Test (SAT) scores in reading and math for fourth graders and Basic Competency Test (BCT) scores in reading and math for third graders, totaling 29,544 third and fourth graders in 690 schools	*School:* ACT composite score (includes English, mathematics, social studies reading, and natural sciences reading), percentage with 5 years experience, percentage with master's degree	*Student:* Race, sex, age *Grade level:* New students as percentage of fourth-grade students *School:* Class size, percentage students receiving free and reduced-price lunch *Zip code (about 3 schools):* Adults' education levels, per capita income (natural logarithm) *District:* Enrollment, public school students as a percentage of all students, percentage of district that is urban, whether district is a city

(*continued*)

APPENDIX (*Continued*)

Study citation	Achievement data	Teacher quality variables	Other controls	Additional features
Goldhaber and Brewer (1997a)	*Student:* Tenth-grade students' test scores from the NELS:88: 5,113 in mathematics, 4,357 in science, 6,196 in English, 2,943 in history	*Teacher:* Years of experience at the secondary level, certified, certified in subject, BA major in subject, MA or more, MA in subject	*Student:* Sex, race/ethnicity, parent education, family structure, family income *Teacher/classroom:* Class size, class percentage minority, teacher race, teacher gender, urbanicity, region, school size, school percentage of teachers with MA or better *School:* Students percentage White, students percentage single parent	
Goldhaber and Brewer (2000)	*Student:* Twelfth-grade students' test scores from the NELS:88: 3,786 in mathematics, 2,524 in science	*Teacher:* Certification in math (standard, probationary subject, private school, none), certification in science (standard, probationary subject, private school, none), years of experience, major in subject, major in education, degree level (BA or less, MA, higher than MA, Ed Spec.)	*Student:* Gender, race, mother's education, father's education, family structure, family size, family income *Teacher/classroom:* Teacher race, teacher gender, class size, class percentage minority *Grade level (12th grade):* Students percentage White *School:* Urbanicity, region, percentage students from single-parent families, percentage students free and reduced-price lunch	
Hanushek (1992)	*Student:* ITBS scores in reading and vocabulary from the Gary Income Maintenance Experiment for 1,920 elementary students	*Teacher:* Score on "Quick Word Test," years of experience (natural logarithm), master's degree	*Student:* Sex, parent income, number of children, permanent income (average 1970–1975), test scores of other family members relative to current child *Teacher/classroom:* Gender, race, class size	

Harnisch (1987)	*Student:* Tests of verbal skills, math, science, and composite of all three from HS&B for 18,684 secondary students at 800+ schools	*School:* Teachers percentage with advanced degree (MA, MS, or PhD)	*School:* SES (index of parents' education, parents' income, father's occupation, household items, parent involvement), students percentage Black or Hispanic, students percentage in academic curriculum, students' self-concept, students' locus of control, students' work orientation, students' family orientation, students' community orientation, school has competency requirement, school size, quality of facilities, student/teacher ratio, teacher turnover rate, number of math and science courses offered, number of other courses offered, total hours in school year, quantity of instruction, disciplinary rules, ability grouping
Kiesling (1984)	*Student:* California Achievement Test scores and New York state CRT scores from 3,374 elementary students in 176 classes	*Teacher:* Years of experience (natural logarithm), degree level (1–9)	*Student:* Father's occupation category *Teacher/classroom:* Class percentage designated as compensatory education students, duration and characteristics of instruction, preparation time, class size
Link and Mulligan (1986)	*Student:* Comprehensive Test of Basic Skills for 2,089 elementary students	*Teacher:* Years of experience (0–5; 6+)	*Student:* Gender, mother's education, father's education, books in the home, family income, home ownership, rooms per family member, hours of instruction in math, hours of instruction in reading, teacher opinion on student need for compensatory education *Teacher/classroom:* Class size

(*continued*)

APPENDIX (*Continued*)

Study citation	Achievement data	Teacher quality variables	Other controls	Additional features
Link and Ratledge (1979)	*Student:* Reading score on California test bureau's Comprehensive Test of Basic Skills for 500 fourth graders	*Teacher:* Degree level (<BS, BS, BS+, MS), age, White, years of experience	*Student:* Sex, race, age, nationality, whether occupation of father is professional, student perception of teacher expectations, student perception of parent expectations, student preference for race of teacher met, student IQ, student motivation *Teacher/classroom:* Number of students in class, racial composition of class	
Maynard and Crawford (1976) (information from Greenwald, Hedges, and Laine, 1994)	*School:* Composite score from 34 schools, including both elementary and secondary	*School:* Teacher education, teacher experience	*School:* Per pupil expenditures, teacher-pupil ratio, school size	
Monk and King (1994)	*Student:* Composite math and composite science test scores from the Longitudinal Study of American Youth (LSAY) for 2,831 students in math and science classes: 977–1,955 students in mathematics; 912–1,936 students in science	*Teacher:* Years of experience, courses in math, courses in life science, courses in physical science Course measures combine graduate and undergraduate courses All variables available for proximate teacher, immediate past teacher, student-specific teachers (past 2 years), and set of all teachers in student's school	*Student:* SES (composite from LSAY data set), course taking by student, whether courses were normal, advanced, or remedial	

Murnane (1975)	*Student:* Standard scores on Metropolitan Achievement Tests of reading and arithmetic for 410–440 students in Grades 2 and 3	*Teacher:* Years of experience (0–2; 3–5; 6+), master's degree, undergraduate major in education, undergraduate GPA, sex, Black, marital status	*Student:* Sex, living in public housing, days present *School:* Student turnover, class size *Community:* Percentage renters on block, percentage students in female-headed families on block
Murnane and Phillips (1981)	*Student:* ITBS scores in vocabulary from the Gary Income Maintenance Experiment for elementary school students in Grades 3, 4, 5, and 6 ($N = 199–277$)	*Teacher:* Years of experience (≤7; 8–14; ≥15), holds master's degree, verbal ability score, attended prestigious college (unspecified), White, female	*Student:* Unspecified demographic characteristics and home environment variables from interviews with parents *School:* Unspecified school characteristics collected from school records
Raymond, Fletcher, and Luque (2001)	*Student:* Texas Assessment of Academic Skills (TAAS) scores in reading and math for students in Grades 3 to 8 in Houston Independent School District ($N = 80,608–132,021$)	*Teacher:* Years of experience (years, years squared), whether hired through Teach for America	*Student:* Race/ethnicity, poverty *Classroom:* Last year's average score of students who were in the same class last year *School:* Percentage African American, percentage Latino/Latina, percentage in poverty
Rivkin, Hanushek, and Kain (2001)	*Student:* Texas Assessment of Academic Skills (TAAS) scores in math for three cohorts of students in Grades 3 to 6, totaling 906,206 students	*Grade and subject:* Percentage of teachers with 0, 1, or more years of experience, percentage with graduate degree	*Student:* Student fixed effects *School and grade:* School-by-grade fixed effects *Grade and subject:* Average class size
Rowan, Chiang, and Miller (1997)	*Student:* High school students' mathematics test scores from the NELS: 88 for 5,381 students	*Teacher:* Major in mathematics (undergraduate or graduate), math test item from NELS teacher survey	*Student:* SES composite, sex, educational expectations, track, course taking, eighth-grade score average of reading, science, and social studies (in addition to mathematics pretest scores described in text)

(*continued*)

APPENDIX (*Continued*)

Study citation	Achievement data	Teacher quality variables	Other controls	Additional features
			Teacher/classroom: Teacher expects the student to go to college, teacher control, teacher reports supportive leadership, teacher reports staff cooperation, teaching emphasis on teaching for higher order thinking, teacher motivation *School:* 15 school-level variables addressing public, size, scheduling, internal structure, leadership, average test score, average SES	
Summers and Wolfe (1975a, 1977)	*Student:* Composite scores on the ITBS for 627 sixth-grade students, 553 eighth-grade students, and 716 twelfth-grade students	*Teacher:* Years of experience, NTE common examination scores (elementary only), NTE area examination scores (secondary only), rating of teacher's college (Gourman rating ≥525)	*Student:* Estimated family income (estimated using census data, knowledge of location, and race), race, sex, unexcused absences and lateness, participated in Head Start, residential moves, born in U.S., early test scores (in addition to pretest scores described in text) *Teacher/classroom:* Class size (≤28; 28–33; ≥33) *Grade level:* Size of pupil's grade *School:* Enrollment, library books per pupil, number of disruptive incidents, peers percentage high achievers, peers percentage Black, physical facilities, characteristics of principal, expenditures on counseling, expenditures on remedial education	

Note. EEOC = Equality of Educational Opportunity Commission; HLM = hierarchical linear modeling; ITBS = Iowa Test of Basic Skills; IV = instrumental variables; HS&B = High School and Beyond; NELS:88 = National Education Longitudinal Study of 1988; SES = socioeconomic status.