The Importance of Methodology in Teasing Out the Effects of School Resources on Student Achievement

CRPE Working Paper 2007-5.0
August 15, 2007

Dan Goldhaber
Center on Reinventing Public Education
Daniel J. Evans School of Public Affairs
University of Washington

The suggested citation for this working paper is:

Goldhaber, Dan. *The Importance of Methodology in Teasing Out the Effects of School Resources on Student Achievement.* Working paper no. 2007-5.0. Center on Reinventing Public Education (CRPE), University of Washington, 2007. Web. <http://www.cedr.us/publications.html>.

I. Estimating the Effect of Educational Resources for Policymaking

Policymakers wishing to make empirically based judgments about which educational resources are worthwhile investments rely on the statistical inferences made by researchers. In the realm of education research, however, few true experiments are designed to discern the *causal* relationships between resources and student outcomes.[1] Instead, most of what we have learned about how these resources influence student achievement has come from studying observed patterns of achievement in schools educating different types of students with varying levels of different educational resources. Unfortunately, educational research studies using non-experimental methods often reach erroneous conclusions because school environments don't in fact resemble experimental conditions and researchers fail to account for this in their research design; often they cannot due to data limitations.

The case of spending on schools is one good example of this. Most studies linking per-pupil expenditure to student achievements find this relationship—if any is found at all—to be quite weak. From this we might conclude that educational spending has little impact on the productivity of schools, which may well be true if, as some have suggested, additional monies tend to be poorly invested (Hanushek, 1996). However, some competing explanations for the weak relationship come to mind. For instance, over time we might see that schools increase spending with no additional output because the mix of students changes, and schools are serving more needy students today than in the past (Berliner and Biddle, 1995). We might also make a snapshot observation that schools are spending money to address student disadvantages. Schools in high-poverty communities may be investing in supplemental tutoring for low-achieving students, or in metal detectors for improved school safety/security. The general point is that the simple *association* between student achievement and spending levels could be misleading if one does not consider mediating factors, such as the existing achievement levels of students.

It is possible to uncover the underlying causal relationship between educational investments and student outcomes with the right data and appropriate statistical techniques. Policymakers and researchers have devoted considerable energy investigating whether certain attributes – credentials, individual characteristics or experiences – can predict teacher effectiveness. This effort continues, and is appropriate given recent

empirical findings identifying teacher quality as the most-important schooling variable influencing student achievement (Rivkin et al., 2005; Rockoff, 2004; Wright et al., 1997). But, as was the case with school spending, it is very possible that the associations we observe between teachers and student achievement might lead us to incorrect conclusions.

Distinguishing between an association (for example, the correlation between two variables) and a causal relationship is more than just an academic exercise—it is crucial to policymakers. We might, for instance, observe that teachers with master's degrees tend to be more effective at raising student achievement than those who lack this credential.[2] Were it a causal relationship arising because teachers acquire key classroom skills as a consequence of going through masters training, one might urge policymakers to require this degree of all teachers. But if this were just an association arising, for instance, because teachers who get their master's degrees tend to be the ones who are very enthusiastic about the profession and also happen to be good teachers, then we would not expect a teacher to be made more effective because she completed a masters program. Thus, requiring this degree of all teachers (a policy which currently exists in some states) would be a waste of educational resources.

The US Department of Education's research arm, the Institute of Education Services, administers the *What Works Clearinghouse* "to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education" (http://www.whatworks.ed.gov/whoweare/overview.html). The evidence standards for scientific research are very high: only randomized controlled trials and strong quasi-experimental designs meeting specific conditions on non-experimental data are considered to meet the evidence standards. Quasi-experimental designs with fewer conditions or randomly controlled trials with problems are categorized as meeting the evidence standard with reservations; and, studies based simply on regression analysis do not meet the standards required to show causality.

The example of master's degrees discussed above is just one of the reasons why we should be cautious about making causal claims based on research in non-experimental settings. Even relatively sophisticated statistical methods (such as multiple regression

models) are prone to misinterpretation, and the results of such models are frequently reported and discussed without the proper caveats.

In this paper, I discuss the role of methodology in assessing the effects of teacher qualifications on teacher effectiveness (also commonly referred to as the "quality of teaching"). I focus on some of the statistical problems that can arise in a non-experimental setting and the ways in which statisticians attempt to deal with them. Then, using data on teachers and students from a statewide administrative dataset from North Carolina, I show how estimates of the impact of various teacher qualifications change depending on the type of statistical methodology applied to the data. The next section briefly reviews the literature on teacher effects and examines how we might think about findings on the impact of teacher qualifications on student achievement in light of the statistical issues can arise.

II. The Quiet Revolution in Methodology

Researchers have been assessing the relationship between teacher qualifications and student achievement ever since the release of the "Coleman Report" (Coleman et al., 1966). Since then, literally hundreds of papers have been published relating various teacher attributes to student achievement (Hanushek, 1997), but mostly focused on two qualifications: degree and experience level. Surprisingly, these studies have yielded a very mixed picture about the relationship between these teacher characteristics and student achievement. For example, in a widely cited synthesis of many of these studies, Hanushek (1986) concludes that "the results are startlingly consistent in finding no strong evidence that … teacher education, or teachers' years of experience have an expected positive effect on student achievement" (p. 1162). But, in a similar review of the literature, Greenwald et al. (1996) conclude: "variables like teacher academic ability, teacher education, and teacher experience show very strong relations with student achievement" (p. 384).

These divergent conclusions stem, at least to some extent, from the choice and weighting of the articles chosen for inclusion (Krueger, 2003), but both reviews include a number of articles that are methodologically flawed. In fact, there has been a revolution in the methodologies used to assess the relationship between teacher qualifications and

student outcomes (usually students' performance on standardized tests), driven in no small part by the availability of new datasets that link individual teachers to their individual students (Brewer and Goldhaber, 1996; Goldhaber and DeArmond, 2007). Simply put, many of the methodological practices that were commonly used in the past are no longer credible today because we know that they produce flawed estimates. The statistical issues associated with these traditional methodologies tend to fall into three general areas: specification issues, measurement error, and omitted variables bias. Complicating matters is the fact that these issues are often interrelated, so that properly accounting for one type of issue may exacerbate problems with another.

Specification issues relate to our understanding and ability (given data constraints) to model student achievement. For example, many older student achievement models, usually referred to as "educational production function" studies, failed to measure the gains in student achievement over a specified period of time, which can be done either by regressing the gain against control variables or by including a measure of initial level of student achievement as a control. In these studies, therefore, it is not possible to determine whether a student's achievement, in say the 10^{th} grade, is a consequence of the quality of schooling in that grade, or of the quality of their elementary schooling experiences.

Sometimes problems arise from the way that variables are specified. For instance, the value of an additional year's worth of experience at the beginning of a teacher's career is unlikely to be the same as the value of an additional year toward the end of her career, since teachers learn many of their classroom management skills in the first couple of years in the classroom.[3] Despite this, the great majority of studies relating teacher experience to student achievement enter experience into a regression as a linear term, effectively treating the value of additional experience as being the same regardless of the point at which it occurs in a teacher's career.

One can also point to deficiencies in the way that researchers have specified teacher degrees. Goldhaber and Brewer (1997a,b), for example, find that a masters degree does not, in general, predict teacher effectiveness, but that degrees in math and science do—for teachers *teaching in those subjects*.[4] Complicating matters is the likelihood that certain teacher credentials may predict effectiveness in some teaching contexts, but not

others. Monk and Rice (1994), for instance, find that the number of math courses taken by teachers while in college predicted 8th grade students' achievement, but only when teachers are assigned to teach the more-advanced math courses.[5] And Eberts and Stone (1985) find little evidence that teachers' mathematics training has any impact on elementary level (4th grade) students' achievement in math.[6]

While we have learned a great deal about specifying achievement models – the move to "value-added" methodology and the recognition of the way teacher experience effects should be modeled are good examples – it is worth noting that, even today, significant uncertainties remain about how to specify student learning gains (Rivkin, 2006; Todd and Wolpin, 2003). For example, the most common approach to measuring achievement involves regressing achievement in year t against individual student and schooling characteristics, including a measure of prior student achievement, usually in the prior year (t-1). But this approach can lead to biased coefficient estimates if the measure of prior achievement is correlated with any unobserved factors influencing achievement in time t, which could arise if students are sorted among teachers based on prior achievement (Rivkin, 2006); this issue also arises below in the context of omitted variables bias. A less-widely employed alternative regresses the growth in achievement (the change from time t and some earlier period) against individual and schooling characteristics. This approach implicitly assumes that there is no decay in an individual's prior learning (Todd and Wolpin, 2003). It is possible to empirically test the appropriateness of a particular approach and correct any model mis-specification, but only with data structures that are typically unavailable.[7] And, unfortunately, it is usually not possible to determine whether estimated teacher coefficients from these models are biased upward or downward, since the direction (and extent) of bias will depend on rates of learning decay and how prior achievement is related to the other factors influencing student achievement.

Measurement error in the context of teacher qualifications often arises in studies where the observational unit being utilized is at an aggregate level—say at the school or district level, as opposed to at the individual teacher level.[8] In this case, the teacher characteristics of interest are not those of individual teachers but rather a school- or district-level variable such as mean years of experience, mean licensure scores, or the

percentage of teachers with at least a masters degree. Measurement error on its own causes estimated coefficients to be biased toward zero (Greene, 2000), implying that aggregate-level studies of teacher qualifications would tend to understate the true predictive power of a teacher holding a particular qualification.

A third major source of biased coefficient estimates is omitted variables bias. This occurs when the achievement model includes at least one explanatory variable that is correlated with a variable that affects student achievement but is excluded from the model. There are several reasons to worry about this source of bias when studying the impact of teacher qualifications on students. First and foremost, teachers are not randomly distributed across students. While it may not happen in all schools or districts, more-credentialed and more-experienced teachers tend to be matched with higher-achieving students (Betts et al., 2000; Kain and Singleton, 1996; Lankford et al., 2002; Loeb, 2001). This is not a problem *per se*. If a model controls for all the factors that influence achievement, then its estimates of teacher qualifications will be unbiased. If, however, (as is almost certain in any statistical model) various influences on student achievement are not fully accounted for in the model (such as parental participation in their children's schooling or peer influences) and achievement influences the student-teacher match, then the coefficient estimates on teacher qualification variables will be biased.

Omitted variables bias likely helps to explain the phenomenon that Hanushek et al. (1996) refer to as "aggregation bias." They show that a systematic pattern exists in the literature on schooling, whereby the more aggregated the data is, the more likely that the analysis will find positive associations between schooling inputs (such as teacher credentials) and student achievement. This is the opposite of what one would expect, given that the measurement error associated with data aggregation should bias coefficient estimates toward zero (as noted above). Hanushek et al. go on to conclude that a likely explanation for this is a correlation between the schooling inputs and omitted school, district, or state factors related to student achievement. For example, we might expect districts that can afford to hire mainly highly credentialed senior teachers to also be likely to have more-affluent parents who contribute to their children's schooling in myriad ways that statistical models do not account for. Similarly, we might expect correlations in state

policies and schooling inputs. States that, for instance, require teachers to attain a masters degree after a set amount of time, may also be more likely to make other investments in children (such health care) that affect achievement.

While aggregation may increase the likelihood that a statistical model suffers from omitted variables bias, it does not follow that disaggregated studies are therefore immune. Many causes of the nonrandom sorting of teachers and students are likely to be related to student achievement, regardless of the level of data under examination. Active parents who have the resources and capacity to support learning in the home are likely to both consider school quality when making residential decisions and put pressure on schools to place their children in classrooms with effective teachers. And, more-experienced and credentialed teachers often have greater say over their school or classroom assignments, either because of explicit contract rules, cultural practices in schools, or labor market bargaining power.[9] In a recent paper, Clotfelter et al. (forthcoming) show that, unless a model explicitly addresses the issue, the nonrandom match of teachers and students is likely to lead to upwardly biased teacher coefficient estimates even when the analysis undertaken is based on a micro-level dataset where teachers are linked to their individual students.

Related to this issue of omitted variable bias is the general problem of self-selection. As described above, teachers (along with employers) play some role in selecting their districts, schools, and classrooms. But teachers also select many of their qualifications, having chosen, for instance: their major and degree level; their performance (e.g. GPA) while in college; a route into the teaching profession; and, by virtue of deciding whether to stay in the profession, their experience level. We can only observe the effects of various qualifications for those who are in the profession and who have chosen to pursue a certain course of action; as a result, any unobserved individual characteristics that lead teachers to pursue that course of action may confound our estimates of the returns to their qualifications. The potential omitted variables in this case are factors such as individual initiative, effort, and commitment to profession.

Omission of these difficult-to-quantify individual characteristics does not necessarily mean that the coefficient estimates for teacher qualifications do not predict teacher effectiveness; rather, it means we should be wary about attaching *causal*

interpretations to these estimates. As mentioned above, one could imagine that teachers with masters degrees or those who have more educational coursework are more effective, but such findings do not necessarily mean that teachers *become* more effective as a consequence of their training. Research on the National Board for Professional Teaching Standards (NBPTS) provides a good example of this. Several studies suggest that National Board Certified Teachers (NBCTs) are more effective than those who do not hold this advanced certification (Cavalluzzo, 2004; Clotfelter et al., 2007; Goldhaber and Anthony, 2007), but some of these same studies that examine teachers before and after attaining this certification do not find any evidence that teachers are made more effective by having completed the NBPTS process.

While less straightforward, the same argument can be made about teacher experience. Teachers choose to enter and to remain in the profession, and research shows that attrition is not random in the sense that those who leave tend to be more academically proficient than those who remain (Hanushek & Pace, 1995; Murnane et al., 1991; Schlechty and Vance 1981). Assume that the more academically proficient a teacher is, the more effective he tends to be. Further assume that one is modeling student achievement as a function of teacher qualifications, but the model fails to fully control for a teacher's academic proficiency (most educational production functions only include such crude measures as degree level and certification status). Such a model will understate the true impact of additional years of experience, because the experience variable will necessarily capture the decline in the academic proficiency of the workforce that tends to be occurring contemporaneously with teachers gaining experience (because the more academically proficient teachers are the ones who leave the workforce so are less likely to show up in the data as more experienced teachers). What these examples illustrate is that the self-selection of teachers into qualifications means that estimated teacher qualification coefficients can sometimes be unbiased predictors of teacher effectiveness, but one should not necessarily treat them as causal.

So how can researchers account for the various issues discussed above? At the most basic level, it is necessary to account for "growth" in student achievement in a value-added framework. This may be done by regressing achievement in year t against various student and schooling variables, including a measure of prior (typically t-1)

student achievement.[10] The argument for doing this, as briefly discussed above, is that, in the absence of measuring growth, one cannot determine whether current student achievement is a result of current educational inputs (e.g. school quality) or inputs from previous years. A student, for example, may do extremely well in middle school grades either because she has excellent middle school teachers or because she had terrific teachers while in elementary school. Some, but certainly not all of the studies used in widely cited meta-analyses (Greenwald et al., 1996; Hanushek, 1986, 1997) meet this basic standard.

But, measuring growth in achievement is likely not good enough if one wishes to accurately identify the impact of schooling attributes on student achievement. The reason is that schools play only a secondary role in determining achievement, which is no surprise given that students typically spend between 15- 20 percent of their time in a school building. Measures of prior achievement are likely to account for only part of the impact of what is happening in a student's home. Thus, researchers typically also include some measure of parental income (e.g. whether a student is eligible for free or reduced price lunch) and/or parental educational level, if possible (Hanushek, 1979).[11]

In general, the richer the set of information on students and their schooling experiences, including details about their teachers, the more likely it is that researchers will avoid problems with omitted variables bias. And, it is possible to test a variety of econometric specifications of the achievement model to determine appropriate statistical fit with datasets that link individual teachers and students and track both over time (Rivkin, 2006; Todd and Wolpin, 2003). But here is where methodology shows itself to be important, for even very rich datasets that include a comprehensive set of *observable* variables are likely to be insufficient because *unobservable* factors likely play an important role in determining student achievement and the match between students and their schools and teachers. There may be a great deal of heterogeneity amongst students or their parents who appear to be the same based on observable characteristics, like family income. For example, parents who are very involved in their children's schooling in ways that are difficult to quantify (such as turning off the television or helping with homework) are also likely to make residential decisions based, in part, on their impressions of districts and schools.

Unfortunately, we often cannot directly measure the variables that we would like to have information on. However, there are several empirical ways to try to account for unobserved influences. One strategy, for example, that may be used to address the potential problem of nonrandom matching of students to schools and teachers is to find samples of students for whom one might reasonably argue the match is actually random.[12] Another is to explicitly model the selection of students into a particular school, type of teacher, or classroom—a method known as propensity score matching. This method starts by modeling the selection of subjects into groups (such as schools or classrooms) and then makes comparisons between different groups by matching the groups determined most likely to be equal. While this non-experimental method does generally provide unbiased estimates when implemented appropriately, two studies examining its use in educational contexts conclude that relying solely on these methods will not always produce results consistent with experimental methods, and should therefore be considered with other empirical evidence when experiments are not feasible (Agodini and Dynarski, 2004; Wilde and Hollister, 2007).

A more common approach is to use a "fixed-effects strategy" to account for potential omitted variables. Such a strategy relies on the notion that one can determine the impacts of variables, such as teacher qualifications, by examining the variation in those qualifications for student observations (which may be individual students observed multiple times) that are thought to share a common value of the omitted variable. For example, if one believes that students in each school are likely to share similar home environments, then it would make sense to try to determine the impact of teachers by focusing on the impacts of variation in teacher qualifications *within* schools. Thus, models that include school fixed effects account for time-invariant school variables, and therefore are likely to address issues of bias associated with the potential nonrandom match of both students and teachers to schools.

Models that include teacher fixed effects identify the impact of teacher qualifications based on variation in credentials within individual teachers—in other words, based on teachers whose qualifications change during the period in which teachers are observed in the data. The potential advantage of using a teacher fixed-effects model approach is that it can help to distinguish between teacher qualifications that *predict*

teacher effectiveness and those that actually *cause* teachers to become more effective. The downside, however, is that while some qualifications may change during the course of a teacher's career – such as pedagogical or content coursework, or experience—in general, teacher fixed-effects models place severe limits on what we can learn about many other teacher qualifications — for example, undergraduate college training or source of entry into the teacher labor market — since these do not vary at all for individual teachers.

Finally, in the case of student fixed-effects models, the impacts of teacher qualifications are identified based on variation within individual students of the qualifications of their teachers — in other words, based on students who are assigned to multiple types of teachers (for instance, more-experienced or less-credentialed) during the period in which they are observed in the data. One might conceptualize this by thinking of students as being on an achievement trajectory, and the identification of the effects of teacher qualifications coming from observing students with particular types of teachers (for instance, a more-experienced teacher) deviate from that trajectory. The advantage of these models is that they account for both *observed* and *unobserved* time-invariant student (and students' family background) heterogeneity.[13] In practice, however, data limitations often preclude estimation of student fixed effects (a minimal requirement is a measure of student achievement in three time periods).

In order to address the multiple issues that can arise, it is necessary to estimate models that include multiple-level effects.[14] However, very few studies simultaneously include school, teacher, and student fixed effects due to the aforementioned data requirements as well as the computational burden involved in doing so.[15] So, how important is it to address omitted variables through a fixed-effects strategy? The next section describes the empirical strategy and data that I use to explore how various model specifications influence the estimates of the effects of several commonly researched teacher qualifications.

III. An Illustrative Example

A. Data and Methodology

To show how changes in model specification impact estimates of several teacher qualification variables, I use data drawn from administrative datasets in North Carolina that are maintained by the North Carolina Education Research Data Center (NCERDC) for the North Carolina Department of Public Instruction of (NCDPI). These records include all teachers and students in the state over a 10-year period (covering school years 1994-95 though 2003-04).[16]

These data are unique and ideal for such a study:they permit the estimation of a number of different student achievement model specifications because students can be linked to their individual teachers (at the elementary level), and both teachers and students can be tracked over time. Furthermore, the North Carolina state assessments are vertically aligned and explicitly designed to permit the estimation of value-added student achievement models.

The student data include individual information about gender, race and ethnicity, disability status, and free or reduced-price lunch status. The teacher data include degree and experience levels, licensure and certification status, the college from which the teacher graduated, and the teacher's performance on one or more licensure exams.[17] For this example, I focus on how model specification affects the estimated effects of four teacher qualifications: whether a teacher has passed the state standard on licensure exams; measures of a teachers experience-level (no experience, 3-5 years of experience, 6-12 years of experience, and 13 or more years of experience); whether a teacher has a masters degree; and whether a teacher is certified by the National Board for Professional Teaching Standards (NBPTS).

Each of these teacher qualification characteristics has important policy relevance. Licensure tests are a key component in determining eligibility to teach.[18] The great majority of school districts award teacher compensation on the basis of degree and experience level (Odden and Kelly, 1997), and there is consensus that teachers become more effective in the earlier years of their career (Goldhaber, forthcoming). And NBPTS certification is a relatively new advanced teaching credential that has become recognized and rewarded in a number of states and localities throughout the country (Goldhaber and

Anthony, 2007; Harris and Sass, 2007).

Table 1 reports simple pair-wise correlations between selected teacher variables and student achievement in reading and math. Not surprisingly, there are positive and statistically significant correlations between each of these teacher qualification variables and student achievement in both subjects. Of course, one cannot conclude whether these patterns reflect the fact that teachers with particular qualifications tend to be paired with students of a particular academic level, or whether it is the teachers themselves that impact the level of student achievement.

Table 1. Correlation Between Selected Teacher Qualifications and Student Achievement

	Reading	Math
Teacher passed licensure standard	0.047 p=0.000	0.053 p=0.000
Teacher's years of experience[19]	0.032 p=0.000	0.031 p=0.000
Teacher has any advanced degree (MA, PhD, etc.)	0.025 p=0.000	0.024 p=0.000
Teacher is NBPTS certified	0.021 p=0.000	0.020 p=0.000

To assess the strength of the relationship between teacher qualifications and student achievement, I utilize the following basic educational production function:

(1) $$A_t = \alpha A_{t-1} + \beta STUDENT + \gamma TEACHER + \delta CLASS$$

The left hand side of the equation (A_t) is the achievement of student *i* in year *t*. The model includes controls for achievement in a prior year, A_{t-1}; a vector of individual student characteristics (gender, race, grade, learning disabilities, parents' education levels, lunch program eligibility, and English ability), STUDENT; a vector of other teacher characteristics (gender, race, measures of college selectivity, licensure status, and graduation from an approved NC teacher certification program), TEACHER, in addition to the teacher variables discussed above; and a vector of classroom variables (class size, year of observation, and the minority composition of the students in the classroom),

CLASS.[20]

The effect of a teacher passing the licensure standard is measured against teachers who have not passed; the effect of teachers having 3-5 years of teaching experience is measured against those who have less than one year of experience; the effect of a teacher having a masters degree is compared to teachers who hold only a baccalaureate degree; and the effect of teachers being NBPTS-certified is measured relative to those who do not hold this certification. The math and reading student achievement measures in the data are normalized within grade and year to have a mean of zero and a standard deviation of one, which means that the coefficients of the teacher qualifications should be interpreted as their predicted impact on a students position in the achievement distribution.

The analyses are restricted to students in grades 4-6 who have a valid math and/or reading "pre-" and "post-" test score (for example, the end-of-year 4th-grade math score would be used as the post-test when a student's end-of-year 3rd-grade math score was used as the pre-test). The major reason for these restrictions is that they allow an analysis for a group of students who are highly likely to be matched to their teachers of math and reading.[21] Further, we restrict the data to those teachers for whom we have valid scores for the Praxis exams, which is required for teacher certification in North Carolina beginning in 1997. This yields a sample of 4.088 unique teachers (9,924 teacher observations) and 177,570 unique students (196,888 teacher-student observations).

As discussed in the previous section, there are several potential sources of bias when estimating the effects of teacher variables. To address these, I exploit the longitudinal nature of the dataset to estimate variants of equation (1) that include either school or student fixed effects (the measure of prior achievement, A_{t-1}, is omitted from the student fixed-effects models), or the two simultaneously. The school fixed-effects models account for time-invariant school characteristics, such as the potential non-random match of teachers to schools; in these models, the teacher qualification coefficients are identified based on within-school variation in teacher characteristics. The student fixed-effects models account for time-invariant student characteristics, such as student motivation or parental support, and identify the teacher qualification coefficients based on variation in the qualifications of individual students' teachers. The models that include both school and student fixed effects account for both school- and student-level

time-invariant unobservables. These models, however, are identified based on variation in the qualifications of individual students' teachers within individual schools, meaning that it is necessary to have a very particular data structure in order to estimate them: students who are linked to teachers and observed with at least 3 teachers (in order to have two achievement gains) within a school.[22]

Finally, I employ a propensity score match (PSM) approach to estimate the relationship between teacher qualifications and student achievement. This is a two-step quasi-experimental design, where I first employ a probit model to estimate the probability that students are assigned to a teacher with a particular qualification, from which the propensity scores are calculated. Next, these scores are used to match student observations so that students whose teachers have a particular qualification are grouped with students whose teachers do not have that qualification. Since students are matched along their estimated propensity scores, the treatment and control groups are comprised of students approximately equal in their likelihood of assignment to a particular teacher, so any differences in the student outcomes are attributable to the differences in the teacher characteristics, and not to the students themselves.[23]

B. Findings

Tables 2 and **3** report the coefficient estimates for the selected teacher qualification variables for each type of model specification, for reading and math models respectively.

Table 2. Estimated Impact of Teacher Qualifications on Students' Reading Achievement

	(1)	(2)	(3)	(4)	(5)
Teacher Characteristics	**Basic Value-Added Model**	**School Fixed Effects**	**Student Fixed Effects**	**School & Student Fixed Effects**	**Propensity Score Match**
Passed Licensure Standard	0.012	0.002	0.030*	0.028	0.006
	(0.012)	(0.007)	(0.014)	(0.015)	(0.011)
Has 3-5 years of experience	0.051**	0.041**	0.013	0.015	--
	(0.007)	(0.005)	(0.011)	(0.012)	--
Holds an advanced degree	0.000	-0.005	-0.001	-0.002	-0.021**
	(0.007)	(0.005)	(0.010)	(0.010)	(0.007)
NBPTS Certified	0.034	0.029*	-0.01	-0.019	0.027
	(0.019)	(0.014)	(0.034)	(0.037)	(0.021)
Robust standard errors in parentheses					
* significant at 5%; ** significant at 1%					

Table 3. Estimated Impact of Teacher Qualifications on Students' Math Achievement

	(1)	(2)	(3)	(4)	(5)
Teacher Characteristics	**Basic Value-Added Model**	**School Fixed Effects**	**Student Fixed Effects**	**School & Student Fixed Effects**	**Propensity Score Match**
Passed Licensure Standard	0.034*	0.023**	0.026*	0.029*	0.027**
	(0.016)	(0.007)	(0.012)	(0.013)	(0.010)
Has 3-5 years of experience	0.063**	0.045**	0.008	0.008	NA
	(0.010)	(0.005)	(0.010)	(0.010)	NA
Holds an advanced degree	-0.005	-0.010*	-0.001	-0.002	-0.018**
	(0.010)	(0.004)	(0.008)	(0.009)	(0.007)
NBPTS Certified	0.020	-0.005	0.029	0.035	0.026
	(0.026)	(0.012)	(0.030)	(0.032)	(0.021)
Robust standard errors in parentheses					
* significant at 5%; ** significant at 1%					

The first thing to note about the results is—adjusting for student and schooling characteristics (in the basic value-added models reported in column 1)—how different they are, in general, from the correlations presented in Table 1. Whereas the correlations suggest a strong relationship between teacher qualifications and student achievement across the board, these value-added models show a statistically significant positive relationship only in specific cases: for teacher experience in reading, and for experience and the licensure standard in math.

Reading left to right across columns 1 through 4 shows the extent to which the addition of school and student fixed effects changes the coefficient estimates, and possibly tends to ameliorate problems with omitted variables. Surprisingly, no consistent pattern emerges. Estimates of the effects of passing the licensure standard and having an advanced degree remain quite consistent across specifications, while the other estimates bounce around more, both in terms of statistical significance and sign. The propensity score match (PSM) models (reported in column 5 of each table) most-closely approximates the school fixed-effects estimates. This suggests that the non-random matching of students to teachers *that one can take account of using observable characteristics* takes place at the school level.

The PSM procedure, while intuitively appealing, can only create comparisons based on observable characteristics, so it does not account for omitted variables.[24] This may be why the most notable changes in the estimated coefficients occur when moving from the school to student fixed-effects specification. The estimates for the returns to experience in particular (and to some extent NBPTS certification) change dramatically—moving from statistically significant and positive (in columns 1 and 2) to insignificant (in columns 3 and 4).

It is conceivable that the estimated positive relationships between the teacher qualification variables and student achievement are simply a byproduct of the match between certain types of students and teachers *within schools* (the school fixed-effects models account for the match of teacher to schools), and that the true relationship (or lackthereof) is unmasked with the addition of student effects. But, there are at least two reasons to be cautious about over-interpreting changes in coefficient estimates and statistical significance. First, in these models, the impact of the teacher credentials are

identified based on only a subset of students in the data: those who have at least two gain scores and have teachers whose qualifications vary. In some cases, this represents a very small number of teachers, so the estimates are more likely to be influenced by outliers in the data. Second, this also means that the coefficients are estimated with less precision. Some of the coefficients, in fact, change very little in magnitude (for example, NBPTS certification in math models), even when they change in their significance level.

IV. Policy Implications and Conclusions

In theory, data analyses can provide information that helps policymakers make good decisions and better allocate scarce resources to enable practitioners to improve student achievement. And, as described in Section II above, advances in data collection and availability have yielded insights into the value of various educational interventions, and established the importance of teacher quality as a determining factor in student learning. But, as the findings presented in Section III illustrate, the estimated impact of various teacher qualifications depends a great deal on the research methodologies that researchers employ. While there is certainly no hard and fast rules that can be used to determine the accuracy of estimated teacher effects in a particular study, it's clear that simply statistical methods, e.g. correlations, tend to over predict the importance of teacher qualifications in determining student achievement.

These findings have important implications for research syntheses. There are enough research studies on teachers that, in doing a review or meta-analyses, one could probably arrive at study inclusion rules that allow for any desired conclusions. The debate over the efficacy of licensure is a good illustration of this. Walsh (2001) in a review of over 200 studies on licensure finds reaches the conclusion that teacher licensure is of little value. Darling-Hammond (2002), citing a somewhat different body of work (there is some overlap in cites, but with different interpretations), reaches the a quite different conclusions that, if anything, teacher licensure policies ought to be more restrictive in governing who is eligible to teach. But, in yet another recent review, Wayne and Youngs (2003) find only two papers on the impacts of teacher licensure that they judged to be rigorous enough (i.e., they were peer-reviewed, used longitudinal

student-level achievement data, and controlled for prior student achievement and student SES) to meet the requirements for inclusion in their review.

It is a judgment call, but, like Wayne and Youngs, I tend to be of the mind that one ought to focus on studies that use more sophisticated econometric approaches in analyzing the effects of teacher qualifications.[25] This means that studies that include, for instance, school or student fixed effects should be weighted far more heavily than studies that only adjust for student covariates, which themselves ought to be weighted more heavily than those that use even simply methods. Not to do this would ignore the fact that the ability to link teachers to their students has led to a quiet revolution in our ability to estimate sophisticated statistical models that can directly address statistical issues, such as the nonrandom matching of students and teachers. Again, it is a judgment call, but I believe that older methodologies for estimating the effects of teacher qualifications (such as simple correlations or value-added models) produce less-than-credible estimates of the value of these qualifications. This is born out by the empirical results in Section III, which show not only that the regression results are quite different, in general, from the simple pair-wise correlations between teacher qualifications and student achievement, but that the direction, magnitude, and statistical significance of the effects can also vary depending on model specification.

Some findings are reasonably consistent across specifications. For instance, the consistent positive findings on teachers passing the licensure tests provide good evidence that this should be considered a predictor of effectiveness. And the consistent lack of any positive findings on teachers having an advanced degree should probably indicate that this qualification has little value. The results for experience and NBPTS certification likely indicate that simple econometric models overstate their importance in predicting teacher effectiveness.

These findings and similar ones reported elsewhere (Clotfelter et al., forthcoming; Goldhaber and Anthony, 2007; Lankford et al., 2006; Rivkin et al., 2005) have important policy implications. While research on schooling has clearly shifted to an emphasis on experimental designs (Goldhaber and Brewer, forthcoming), it is likely that most educational research will nevertheless continue to be non-experimental. Given the considerable variation within schools in the allocation of resources, and the fact that

students are most likely not randomly matched to teachers, analyses using school aggregates can be problematic, which puts a premium on having data that allow for the type of analyses discussed in this paper. It is only recently (since the mid-1990s) that any state or national database has allowed for the estimation of models that avoid statistical problems created when only school-level information is available, and this information is currently available in only a handful of states. Hopefully this will change as policymakers begin to see the utility of such data for informing sound public policy.

Acknowledgements –The research presented here is based primarily on confidential data from the North Carolina Education Research Center (NCERDC) at Duke University, directed by Elizabeth Glennie and supported by the Spencer Foundation. This research has been supported by grants from the U.S. Department of Education, Office of Education Research and Improvement (OERI #R305T010084) and National Science Foundation (NSF #REC-0335656). The author wishes to acknowledge the North Carolina Department of Public Instruction for its role in collecting this information, and to thank Mike Hansen for research assistance, Carol Wallace for editorial assistance, and Mary Kennedy for thoughtful suggested revisions on an earlier version of this paper. The views expressed in this paper do not necessarily reflect those of the University of Washington, the Urban Institute, or the study's sponsors. Responsibility for any and all errors rests solely with the author.

References

Agodini, Roberto, and Mark Dynarski. 2004. Are Experiments the Only Option? A Look at Dropout Prevention Programs. *Review of Economics and Statistics* 86 (1):180-194.

Ballou, D. (2005). "Value-Added Assessment: Controlling for Context with Misspecified Models." Paper presented at the Urban Institute Longitudinal Data Conference, March 2005.

Ballou, Dale, William L. Sanders, and Paul Wright. 2004. Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Studies* 29 (1):37-65.

Berliner, David C., and Bruce J. Biddle. 1995. *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Redding, MA: Addison-Wesley Publishing Company.

Betts, Julian R, Kim S. Rueben, and Anne Danenberg. 2000. *Equal resources, equal outcomes? The distribution of school resources and student achievement in California*. San Francisco, CA: Public Policy Institute of California.

Boyd, Donald J., Pam Grossman, Hamilton Lankford, Susanna Loeb, and Jim Wyckoff. 2006. How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement. *Education Finance and Policy* 1 (2):176-216.

Brewer, Dominic J., and Dan D. Goldhaber. 1996. Educational Achievement and Teacher Qualifications: New Evidence from Microlevel Data. In *Optimizing Education Resources*, edited by B. S. Cooper and S. T. Speakman. Greenwich, CT: JAI Press.

Cavalluzzo, Linda C. 2006. *Is National Board Certification an Effective Signal of Teacher Quality?* The CNA Corporation 2004 [cited January 13 2006].

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. How and Why do Teacher Credentials Matter for Student Achievement?: National Bureau of Economic Research, Cambridge, MA.

———. forthcoming. Teacher Sorting, Teacher Shopping, and the Assessment of Teacher Effectiveness. *Journal of Human Resources*.

Darling-Hammond, Linda. 2002. Research and Rhetoric on Teacher Certification:

A Response to "Teacher Certification Reconsidered". *Education Policy Analysis Archives* 10 (36).

Decker, Paul T., Daniel P. Mayer, and Steven Glazerman. 2004. The Effects of Teach for America on Students: Findings from a National Evaluation: Mathematica Policy Research.

Ehrenberg, Ronald G., Dominic J. Brewer, Adam Gamoran, and J. D. Willms. 2001. Does class size matter. *Scientific American* 285 (5):79-85.

Ferguson, Ronald F. 1998. Can Schools narrow the Black-White Test Score Gap. In *The Black-White Test Score Gap*, edited by C. Jencks and M. Phillips. Washington, DC: The Brookings Institution.

Goldhaber, Dan. Forthcoming. Teachers Matter, But Effective Teacher Quality Policies are Elusive: Hints from Research for Creating a More Productive Teacher Workforce. In *Handbook of Research in Education Finance and Policy*, edited by b. H. F. Ladd and E. B. Fiske.

Goldhaber, Dan, and Emily Anthony. 2007. Can Teacher Quality be Effectively Assessed? National Board Certification as a Signal of Effective Teaching. *Review of Economics and Statistics* 89 (1):134-150.

Goldhaber, Dan, and Dominic Brewer. Forthcoming. What Gets Studied and Why: Examining the Incentives that Drive Education Research. In *The Politics of Knowledge*, edited by F. M. Hess.

Goldhaber, Dan D., and Dominic J. Brewer. 1997. Evaluating the Effect of Teacher Degree Level on Educational Performance. In *Developments in School Finance 1996*, edited by J. William Fowler. Washington, DC: National Center for Education Statistics.

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

Goldhaber, Dan, and Michael DeArmond. 2007. Teacher Characteristics, Workforce Policies, and the Search for Teacher Quality: Implications for Research and Data. In *Teacher Supply-Demand Symposium for the National Center for Education Statistics*. Washington, DC: American Institutes of Research.

Greene, William H. 2000. *Econometric Analysis*. Fourth Edition ed. Upper Saddle River, NJ: Prentice Hall.

Greenwald, Rob, Larry Hedges, and Richard Laine. 1996. The Effect of School Resources on Student Achievement. *Review of Educational Research* 66 (3):361-396.

Hanushek, Eric. 1979. Conceptual and Empirical issues in the Estimation of Education Production Functions. *Journal of Human Resources* 14 (3):351-388.

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

———. 1997. Assessing the Effects of School Resources on Student Performance: An Update. *Educational Evaluation and Policy Analysis* 19 (2):141-164.

———. 1999. Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis* 21 (2):143-163.

Hanushek, Eric A., and Richard R. Pace. 1995. Who Chooses to Teach (and Why)? *Economics of Education Review* 14 (2):101-17.

Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor. 1996. Aggregation and the Estimated Effects of School Resources. *Review of Economics and Statistics* 78 (4):611-627.

Kain, John F., and Kraig Singleton. 1996. Equality of Educational Opportunity Revisited. *New England Economic Review* (5-6):87.

Kennedy, Mary. 2007. Defining a Literature. *Educational Researcher* 36 (3):139-147.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. 2002. Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis* 24 (1):37 - 62.

Loeb, Susanna. 2001. Teacher quality: Its enhancement and potential for improving pupil achievement. In *Improving Educational Productivity*, edited by D. Monk, H. Walberg and M. Wang. Greenwich, CT: Information Age Publishing.

Mayer, Susan E. 1997. *What money can't buy: family income and children's life chances*. Cambridge, MA: Harvard University Press,.

Monk, David, and Jennifer King. 1994. Multi-level Teacher Resource Effects on Pupil Performance in Secondary Mathematics and Science: The role of teacher subject matter

preparation. In *Contemporary Policy Issues: Choices and Consequences in Education*, edited by R. G. Ehrenberg. Ithaca, NY: ILR Press.
Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Balinger Publishing Company.
Murnane, Richard J., Judith D. Singer, John B. Willett, James J. Kemple, and Randall J. Olsen. 1991. *Who will teach? Policies that matter*. Cambridge, MA: Harvard University Press.
Rivkin, Steven. 2006. Cumulative Nature of Learning and Specification Bias in Education Research. Working Paper. Amherst, MA: Amherst College.
Rivkin, Steven, Eric A. Hanushek, and John F. Kain. 2005. Teachers, Schools and Academic Achievement. *Econometrica* 73 (2):417-458.
Rockoff, Jonah E. 2004. The Impact of Individual teachers on Students' Achievement: Evidence from Panel Data. *American Economic Review* 94 (2):247-252.
Schlechty, P. C., and V. S. Vance. 1981. Do Academically Able Teachers Leave Education - the North-Carolina Case. *Phi Delta Kappan* 63 (2):106-112.
Todd, Petra E., and Kenneth I. Wolpin. 2003. On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal* 113 (485):F3-F33.
Walsh, Kate. 2001. Teacher Certification Reconsidered: Stumbling for Quality. Baltimore: Abell Foundation.
Wayne, Andrew J., and Peter Youngs. 2003. Teacher characteristics and student achievement gains: A review. *Review of Educational Research* 73 (1):89-122.
Wilde, Elizabeth Ty, and Robinson Hollister. 2007. How Close is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment. *Journal of Policy Analysis and Management* 26 (3):455-477.

Endnotes

[1] Examples of experimental design in education research include the Perry Preschool experiment (Ferguson, 1998), Tennessee's Student-Teacher Achievement Ratio (STAR) project (Hanushek, 1999; Ehrenberg et al., 2001), and the Mathematica study of Teach For America (TFA) teachers (Decker et al., 2004).

[2] I use this merely as an example, as significant numbers of research studies suggest that master's degrees typically do not predict teacher effectiveness (Goldhaber and Brewer, 1997a; Hanushek, 1986, 1997).

[3] Studies that do allow for a non-linear relationship between teacher experience and student achievement provide evidence that experience in the classroom matters more in the early years of a teacher's career (Clotfelter et al., forthcoming; Goldhaber and Anthony, 2007; Kain, 1995; Murnane, 1975; Rivkin et al., 2004; Rockoff, 2004).

[4]It is useful to note that selection bias should be accounted for when considering the effectiveness of subject-matter training. If the best teachers in a given subject leave the profession, then subject-matter training will likely appear less effective.

[5] At the elementary level, by contrast,

[6] Of course it is also possible that, in fact likely, that the value of degrees or coursework might vary depending on the institution that a teacher attends and/or the time frame in which the degree or coursework were completed.

[7] For example, one can correct for the correlation between prior achievement and unobserved characteristics by instrumenting for the measure of prior achievement; this approach, however, requires at least three years of student achievement information.

[8] It is also worth noting that, since teachers are clustered into schools and districts, it is important to take account of the hierarchical nature of the data by using a statistical procedure (e.g. Hierarchical Linear Modeling, or error clustering) that accounts for this data structure. Failure to do this will not lead to biased coefficient estimates, but will lead to incorrect standard errors.

[9] This is not terribly surprising, since typically financial rewards are not provided for the difficulty of job assignment. In the absence of such rewards, we might expect teachers to gravitate toward more-desirable positions if they can .

[10] Or alternatively, treating the growth from year t-1 to year 1 as the dependent variable.

[11] As Mayer (1997) illustrates, one also needs to be cautious about interpreting the impact of variables like family income as being causal.

[12] For example, Clotfelter et al., (2006) analyze fifth-grade students in North Carolina by performing tests of significant differences between aggregate classroom characteristics and those of the school to detect non-random assignments of students to teachers. With these tests, the authors identify schools that appear to have random assignment rules and estimate the marginal impact of teacher characteristics on student learning using only this subset of classrooms to produce unbiased estimates.

[13] However, as Ballou (2005), Ballou et al. (2004), and Rivkin (2006) illustrate, there are strong assumptions implicit in these models, and thus some uncertainties about how to interpret their results.

[14] Models that include both school and student fixed effects are typically referred to as "spell" models.

[15] See Harris and Sass (2006) as an example of a paper that does estimate models with both school and student fixed effects.
[16] Student information for 4th and 5th graders for 1996-97 is missing from the dataset.
[17] For more detail on this North Carolina dataset, see Clotfelter et al. (forthcoming) or Goldhaber (forthcoming).
[18] In North Carolina, elementary school teachers are required to take and pass two licensure tests (the Praxis II Curriculum, Assessment, and Instruction test, and the Praxis II Content Area Exercises test) and achieve a combined cut score of 313 to be eligible to receive a full teaching license in the state. This standard was adopted in 2000. Teachers may teach on a temporary licensure for a year without having achieved the licensure test standard. For more on this issue, see Goldhaber (forthcoming).
[19] Note that here I use a continuous measure of teacher experience, whereas I use categorical measures of experience in the statistical models discussed below.
[20] Analyses of the value-added of various school and teacher effects is generally based on one of three empirical specifications: one, like that specified in equation 1, where the dependent variable, a measure of achievement, is regressed against a set of controls that includes a measure of prior achievement; a second, where the dependent variable is the gain (the difference between a post-test and some measure of prior achievement) in test scores regressed against a set of controls; and finally, a specification where achievement is regressed against a set of controls that includes student fixed-effects. I have experimented with all three specifications, and the findings for the teacher qualification variables upon which I focus are qualitatively the same across all three specifications.
[21] Teachers and students are matched based on the teacher of record listed on a students state test. Students are not tested prior to the 3rd grade and they typically switch teachers for grades above 6th.
[22] In the student fixed effects models only 19,150 unique students (approximately 10% of all observations) meet the data structure described and they provide the variation to identify the model here. In the student-school interacted fixed effects models only 16,734 (approximately 8% of all observations) meet the identification criterion described.
[23] The method used to carry out the match in this procedure is 10 nearest neighbors matching, which takes an average of the 10 nearest (in terms of propensity score) counterfactual observations in the dataset to provide a counterfactual estimate. The results were qualitatively similar when either nearest neighbor matching or radial matching were employed instead.
[24] It was not appropriate to use a PSM approach to estimate the impact of experience because it is not a binary variable.
[25] An irony, pointed out by Kennedy (2007), is that some of these more sophisticated research papers may be excluded from some reviews because they don't present information in ways that allow them to be compiled with many older, and far less methodologically sophisticated, papers.