Key Attributes of State Data Systems That Allow for Sophisticated Research on Teachers

Dan Goldhaber

Prepared for the
Center for Teaching Quality
Fall 2005

The Potential of State Databases

Analyses of state-level data can provide information that helps policymakers make good decisions and better allocate scarce resources to enable practitioners to improve student achievement. The ability to link individual students to their specific teachers has led to numerous insights into the value of various educational interventions, and revealed the key importance of teacher quality as a determining factor in student learning.

It is only relatively recently (since the mid-1990s) that any state or national database has permitted the linking of individual students and teachers and the tracking of both over time. However, such a data structure is necessary to avoid statistical problems that can occur when only school-level information is available. For example, many of the older "educational production function" studies (see Hanushek, 1986, for an overview of many such studies) were done linking school-level measures of student achievement (e.g., student test scores averaged to the school level) and school-level measures of school resources (e.g., the percentage of teachers with master's degrees). 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.[1]

In the time that longitudinal, linked student-teacher data has been available, there have been many examples of how it has led to advancements in our understanding of schools. Studies using state data from Texas (Rivkin et al., forthcoming) and Tennessee (Sanders et al., 1997) have shown the overall importance of teacher quality, which has in turn helped the nation to focus more acutely on investigating policy options for improving teacher quality. The Fordham Foundation, the National Commission on Teaching and America's Future (NCTAF) and the Education Roundtable have all recently addressed policy issues regarding teacher quality. For example, NCTAF's second major goal is to "assure quality teacher preparation", and The Fordham Foundation seeks "a solid core curriculum taught by knowledgeable, expert instructors" (NCTAF, 2004; Fordham Foundation, 2004).

Linked student-teacher data is also crucial for determining the value of particular teacher credentials. An example of this can be found in some work I completed with Dominic Brewer (Goldhaber and Brewer, 1997, Goldhaber et al., 1999), which illustrates that aggregate-level data may mask the importance of particular teacher credentials, in this case teacher degree level. Many studies find a weak relationship between teachers holding an advanced (masters) degree and student achievement, a result we replicated when we treated the masters as a generic degree. However, when we took advantage of disaggregated data to investigate whether degree level might matter in some contexts, we discovered that subject-specific teacher background in mathematics and science is systematically related to student achievement in these subjects, even though teachers' advanced degrees in general are not.

[1] For more information on these issues, see Goldhaber and Brewer (1997) and Hanushek et al. (1996)

Longitudinal student-teacher data also permit the estimation of sophisticated statistical models that avoid matching problems that may have plagued earlier educational production function studies. In particular, when analyzing non-experimental data, researchers often worry that the relationship between two (or more) variables thought to be causal is in fact simply a statistical artifact. For instance, were it the case that struggling students were assigned primarily to the most senior teachers, we would likely observe a negative relationship between teacher experience and student achievement that had no causal implications for the effectiveness of teachers as they gained experience. In this hypothetical case, the "bias" in the estimated impact of teacher experience on student achievement results from the fact that the non-random matching of students and teachers has not been properly accounted for. In practice, students who are struggling tend to be assigned to less-experienced and less-credentialed teachers, which leads to an overestimate of the impacts of teacher credentials in conventional statistical analyses. This is aptly illustrated by new research from Clotfelter, Ladd, and Vigdor (2003), who test whether unobservable characteristics cause systematic bias in the estimated effects of observable variables.

The importance of high-quality state data is aligned with the desire to push accountability models from the school level (as they currently are under NCLB) to the teacher level. Some states, (Tennessee for example) already evaluate the value-added of particular teachers, and policymakers appear to be increasingly interested in this pursuit (Gregorian, 2004), for good reason.[2] As I have just described, new research has shown just how important teachers are in determining student outcomes and that there is generally a great deal of variation of teacher quality within schools. These facts at least suggest that teachers ought to be a focus of reform efforts. Furthermore, it is crucial to understand what makes one teacher more effective than another if one wishes to replicate and spread effective teacher practices. Of course, this can only occur if it is possible to accurately identify the relative effectiveness of teachers.

Unfortunately, at this point there are few states that have data systems in place that allow for the type of research and analyses described above. In fact, I am aware of only four states – Tennessee, Florida, Texas, and North Carolina – that have the necessary data elements to longitudinally link students and teachers, and as I detail below, the ability to do this linkage does not necessarily imply that it can be done with ease.

This brief will provide a simple roadmap for states wishing to develop a data structure that permits the estimation of value-added teacher quality models. Specifically, the essential elements of such a data structure will be outlined in the context of describing how I (along with my co-author Emily Anthony) used a comprehensive data set from North Carolina to evaluate the relationship between the certification of teachers by the

[2] The Tennessee Value-Added Assessment System (TVAAS) ranks schools by how their students score on standardized tests; it provides information to school administrators, teachers, parents and the public on how schools are doing in helping students make yearly academic gains. TVAAS tests each student in each grade in a number of subjects, then ranks teachers based on their students' yearly test score gains (TVAAS, 2004).

National Board for Professional Teaching Standards (NBPTS) and elementary level student achievement.[3]

The Key Elements for Estimating Value-Added Models

The key to estimating value-added models is the ability to isolate the impact of various educational variables from other factors that influence student achievement. I believe that, at a minimum, this requires information on student achievement over multiple years, preferably annually in subsequent grades; student background information such as measures of socio-economic status (SES), and learning and English proficiency status; and the ability to link schooling resources (e.g., teacher characteristics, class size, etc.) to *individual* students – so that researchers know, for instance, students' actual class sizes rather than just the average student-teacher ratio in a school – in the years for which the achievement data are available.[4] Additional types of data, described below, permit greater inference and certainty about the estimated impacts of various schooling resources on student achievement.

North Carolina has all the essential data structures that permit the estimation of sophisticated value-added student achievement models. This dates back to the implementation of the "ABC" education reform program (the state's accountability system) in 1996-97, which required the state to begin tracking detailed student achievement data used to evaluate school performance. Specifically, the state designed tests to measure subject objectives defined in the *North Carolina Standard Course of Study*. Student performance on these tests are used by the accountability department of the North Carolina Department of Public Instruction (NCDPI) to determine performance and growth/gain goals and ratings for all schools in the state. Schools' performance levels are determined by the percent of students who obtained test scores that fall into the determined score range for the grade and subject test to achieve a at or above Level III (at grade-level) test score. School growth goals are determined by three factors: first, the state average rate of test score growth in that grade and subject in the second year of state end-of-grade testing; second, an estimate for "true proficiency" of students in a school[5] and third, an estimate for regression to the mean.[6] I discuss North Carolina's ability to measure annual growth in test scores below.

These student tests represent the first key element of the North Carolina dataset. Unlike the majority of states, in North Carolina students are tested annually in grades 3 through 10 using tests that are "vertically aligned," meaning that they are specifically designed to determine individual student achievement *growth* in performance by subtracting the previous year's end-of-grade test from the subsequent year's end-of-grade

[3] This study can be downloaded from www.crpe.org.
[4] For a more comprehensive discussion of these issues, see Hanushek (1986) or Goldhaber and Brewer (1997).
[5] This is measured by subtracting a school's overall combined reading and math scores from the 1994-95 North Carolina overall combined reading and math scores.
[6] This is measured by subtracting: 1.) a school's overall math scores from the 1994-95 North Carolina overall math scores and 2.) a school's overall reading scores from the 1994-95 North Carolina overall reading scores.

test in each subject.[7,8] Vertical alignment is important so that researchers have an appropriate base (a "pre-score") by which to judge student learning gains (the differential between a pre- and post-score).

It is possible to use non-vertically aligned tests as either the "pre" or "post" test score. For example one could use a commonly administered nationally normed referenced test (NRT) test such as the Stanford 9 as the pre- and a state-specific criterion referenced test (CRT) as the post-test. The problem with this is that it is not clear the degree to which the base and follow-up tests are covering the same content material. This means that estimated learning gains may not accurately reflect student learning on particular educational content material.

Another key issue in regards to testing is the scaling of tests. Tests that are not scaled correctly may not consistently measure learning across the test performance distribution meaning, for instance, that a measured test gain for students scoring near the bottom of the test scale may not be equivalent in learning terms to an equivalently sized test gain for students scoring near the top of the scale. At the very top and bottom of the test scales, the tests may suffer from "floor" or "ceiling" effects. A floor effect occurs when a test cannot adequately measure the performance of students who do quite poorly or a test because there is a minimum performance measure of the test (e.g. zero), and a ceiling effect is the same problem at the top end of the scale, when a test cannot adequately measure the performance of students who do extremely well because there is a maximum measured performance.[9]

Issues of test scaling can be controversial in education research. For example, researchers have drawn very different conclusions about the performance of students in Texas, in part at least based on their assessments of the test previously used by the state to measure student learning (the Texas Assessment of Academic Skills, commonly referred to as the TAAS). Some researchers feel the test is a good benchmark for measuring student achievement gains, while others feel it is only a very "low level" test implying it captures the degree to which students master basic material, but does not accurately identify students' learning of more advanced material.

A second key aspect of the North Carolina data is that, in addition to student achievement information, the state maintains a fairly standard set of student background information such as race/ethnicity, gender, learning disability, free or reduced-price lunch status, and English proficiency status. This information is important if one wishes to statistically adjust statistical models for the type of students being taught. For example, students with limited English proficiency often perform less well on standardized tests than do English-proficient students. These students, however, may tend to have larger

---

[7] For more information on how various growth/gain goals and measures are determined by the state, see http://www.ncpublicschools.org/accountability/reporting/ (12/06/04).

[8] Students are actually tested twice in the 3rd grade, which is the first year of student testing, so that one can calculate 3rd grade growth in achievement.

[9] Readers wishing more information on ways to avoid floor and ceiling testing effects should look to research on "Item Response Theory" (IRT) testing.

learning gains. To adjust for these types of differences, researchers must have information on students' backgrounds.

North Carolina also maintains fairly detailed information on its teachers, including standard teacher background information (e.g., race/ethnicity); credentials (e.g., licensure status and area, degree level) and experience; and--unlike many states--information on if teachers attended a state-approved teacher education program and their performance on various tests (e.g., licensure tests such as Praxis I or II, or college entrance exams such as the ACT or SAT). While the detail of the teacher data is important, the unique aspect of the North Carolina data is that it is possible though not necessarily easy, to link students to their individual teachers. This is the third key element that allows for comprehensive research on teacher effects; a teacher's name is associated with each student's test record, which allows for the match of students to their teachers and to the characteristics of their classes.[10] To my knowledge, the only states where this individual student-teacher linkage can be done statewide are: Florida, North Carolina, and Tennessee.[11]

The final key aspect of the North Carolina data is that students and teachers can be tracked over multiple years, in theory for as long as they remain in a public school system in North Carolina. Researchers need at least three years of student-teacher matches in order to estimate the type of statistical models ("student fixed-effects models") that account for the likely non-random match between teachers and students that arises from classroom assignment.[12]

The North Carolina data are, taken as a whole, unique, and to my mind it currently represent the best state data for analyzing various assessments of teacher effects. That said, the data are far from perfect, as the description of our study on the relative effectiveness of NBPTS certified teachers will amply illustrate.

How We Built Our North Carolina Dataset

Because North Carolina has all the key structures permitting value-added analyses of student achievement, as well as a large sample of NBPTS certified teachers, it is an ideal state for studying the effects of NBPTS certification. This large sample gives researchers the power to detect relatively small differences between various NBPTS groups--in this case students taught by teachers who were either certified, non-certified applicants, or non-applicants.

---

[10] The fact that each student can be linked to his/her teacher effectively provides information on that teacher's class, since the class is ultimately the aggregation of all of a teacher's students.

[11] Other state, Texas for instance, are close to this, allowing the linkage of teachers to particular schools and grades.

[12] For example, we might imagine that parents who provide a great deal of educational support in the home would also try to get their students into classes with experienced teachers. We would not want to mis-attribute the impact of parental support to teacher experience. For more information on these type of models, see Goldhaber and Brewer, 1997.

Our study began in 2000, and at that time the state had both student and teacher data available for three school years (1996-97, 1997-98, and 1998-99). Our primary challenge in building a dataset for analyses was to merge teacher and student records. While teacher names are listed on each student's testing record, this does not necessarily mean that the listed teacher taught the student.

After consultation with state and local officials, we opted to limit the focus of the study to elementary schools. The reason is that, in higher grades where students typically switch classes throughout the data, the teacher of record on each student's test is not likely to be the teacher who taught that student. Even at the elementary level, however, we were not sure that the teacher listed on each student's record was the teacher who taught that student. To check on this we surveyed state officials, who stated that at least 90 percent of the time in "regular" elementary schools, the students' classroom teacher is the same person as the one listed on their student records as their "test administrator." The major exception to this rule is "magnet" schools where students are more likely to have different teachers who specialize in particular subjects. Thus, to minimize any mismatching of student and teacher records, we excluded magnet schools from our sample.

We followed up on that information by surveying district-level testing officials about the teacher listed on the student record; they confirmed that at least 90 percent of the time at the elementary level the testing administrators were indeed also students' classroom teachers. We gathered our sample by calling the first 15 districts (in alphabetical order) as well as the five districts serving the largest metropolitan areas in the state, until we received ten districts' responses on the question. Exceptions were said to occur only in rare cases; for instance, when a teacher is absent on a testing day the listed teacher will not be the teacher who taught the students. Thus, we are reasonably confident that by restricting the analyses to non-magnet elementary schools we ensured that an overwhelming majority of the potential student-teacher matches would be correct. Thus, this restriction gave us the parameters of our sample: students in the 3rd, 4th, and 5th grades (recall that the state does not test students before the 3rd grade) from 1996-97 to 1998-99.

While we are "reasonably" confident (and our statistical findings on the effects of NBPTS and other teacher credentials back the proposition that our match was successful),[13] it certainly would be preferable to have greater certainty about this linkage. This could be handled by having the state make it explicit that such a match is required by districts (who are responsible for this type of data), and by including a field in the data reporting form that asked local officials to confirm whether the teacher(s) listed on each student's form is in fact the teacher who taught the student in a particular subject. This would clearly require additional data fields for cases where students are taught by multiple teachers. However, in the absence of this type of information it is not possible to conduct research that requires subject-level matches, for example on the effectiveness of math teachers at the secondary level.

[13] Were the match totally unsuccessful we would expect to find no statistically significant teacher effects as there would be extremely high "measurement error."

The next step was to actually link students to teachers, which proved more difficult than anticipated. Though local districts list a teacher's name on each student's record, not all use the same format for the teacher's names. For example, some list first, middle, then last name; while others list last name, then first name; some list last name with only the first initial of the first name; and some list only last name. We created programs that accounted for the various district formats for teacher names on the student records to link teachers and students for each of the 118 districts in the state, and then used teacher records to determine which teachers had a unique last name/school number combination. For those that did, we linked students and teachers using teacher last name within school. If the teachers' last name/school number combination was not unique, then we linked teachers to students' records using teacher last name/first name initial/school number, or last name/first name/school number combination, depending on which combination was unique within the school. In cases of doubt or insufficient information, we did not attempt to match teachers to students. Besides the varying district information for teachers found on the student records, we also encountered numerous cases with errors (typically occurring when information is electronically coded, for instance, mis-keying a particular character) that prevented some teacher and student matches. Thus, while the great majority of our student-teacher matches were completed using computer programs, a small--but significant--minority of the matches were completed manually by "eyeballing" the records. For example, a computer program cannot easily be programmed to recognize that "jon a smith" should be linked to "Smith, Jonathon A", even though it may be clear they should be linked when the records are examined by the human eye. Overall our linkage of students and teachers was very successful. We were able to match 771,537 of 889,655 student observations (students could appear in our dataset multiple times since we were matching across multiple years) to 32,399 teachers, which is about an 80 percent match rate.

Though there is always the possibility of miscoding, regardless of whether the student-teacher linkage is through teacher name or a numeric code, a numeric code is probably preferable given that it would reduce district confusion over the appropriate way to code the information. One possibility is that states use social security numbers as the linkage identifier. The problem with this is that it allows researchers to identify individual teachers. This creates a dilemma for states who might wish to ensure teacher confidentiality, and may actually make it more difficult for researchers since institutional review boards, responsible for ensuring the protection of Human Subjects, generally also prefer that researchers not have access to individually identifiable information. Thus, an ideal database would require states to have an established uniform teacher identification number that is linkable to both teacher information and student test results.

The next step in building the database was to link students over time. This proved to be far easier, as students were linked based on social security numbers. However, not all students appear in the data year after year. Of the 771,537 students who were matched to at least one teacher, just over 600,000 student-teacher matches appeared in more than one year of our data and were found to have valid scores for the reading and math pre- and end-of-year tests (of course, both scores are needed to measure the gain in student achievement).[14] Unfortunately, for the most part we have no information on why students who are in our data in one year are not in the next (with the exception of those students who are in the 5th grade in any year and then move into 6th, a grade that falls outside of the focus of the study). It is certainly plausible that the students who are lost from year to year simply represent those that moved out of the state, however, we have no way of knowing this for sure. The ability to verify why a student disappeared from the sample would help ensure that there is nothing systematic about those students for whom we lack complete pre- and post-test information.

In the course of constructing a useable dataset, researchers make many decisions about how to code variables and account for inaccurate or missing information.[15] In addition to the missing students discussed above, there are always cases where particular variables are missing or inaccurate for particular observations, and these are standard problems that statisticians face when analyzing data. For example, we might have a teacher for whom we have information on everything but their experience level. In general, the more missing information on a particular variable, the less accurately one would be able to measure the impact of that variable. Furthermore, one might always worry about the quality of the information that is available and whether there is something systematic about the availability of particular information. As an example, researchers investigating the impact of student drug use on academic achievement might worry that self-reports of drug use do not accurately reflect actual use, and that students who do not respond to this question on a survey might be particularly heavy users.

There is no absolute standard for how much missing (teacher or student) information is "too much." In general, the more complete the information, the more precisely one can estimate the effects of teachers or schools on student achievement. Furthermore, there is less concern about whether the data that are missing are systematic in a way that would lead to biased estimates of the effects of schooling variables on student achievement. It is possible to estimate, based on a number of assumptions, the size of the teacher (or other schooling) effects that can be detected given the sample size of the data. It is also possible, again given a number of assumptions, to determine the potential for bias as a result of missing information, but, unfortunately, there is no way to definitively determine whether bias exists as a result of missing data.

While there were cases of missing or inaccurate information in the North Carolina dataset, for the most part the states' records were remarkably complete. This is not surprising given that many of the variables that we were interested in for this study were used by the state for other purposes, such as evaluating schools or determining teaching eligibility or compensation. There are a number of ways to account for cases of missing or inaccurate information. Here I discuss some problems we had with teacher performance on standardized tests, such as licensure exams, to illustrate how we handled the issue of missing or inaccurate teacher licensure test information. This variable was particularly important for our study because we wanted to know not just whether NBPTS

[14] The exception is the 3rd grade where students are tested at the beginning and end of the school-year.
[15] An example of inaccurate information might be if a teacher is reported to have 115 years of experience.

certification represents a good indicator of teacher quality but whether it provides teacher quality information above and beyond that derived from teacher test performance.[16]

There were three problems with the teacher test variables: first, not all teachers had a test score attached to his or her record; second, some coded test information fell outside of what is possible for a particular test (e.g., an SAT score of over 1600); third, not all teachers had the same test information. Dealing with the first two problems was relatively straightforward. We opted to retain only teacher test scores that fell into the proper range for each particular test (e.g., if the range of possible scores for a test was between 100 and 200, and the recorded score was a 54, we considered the test to be missing). We replaced missing teacher test scores with the mean value for that particular test in order to keep as many observations in our models as possible.[17] Close to 90 percent of our student-teacher matches (about 540,000 out of about 600,000) had teachers with at least one valid test score, thus we only had to replace about 10 percent of the sample for this particular variable.[18]

The North Carolina data contains information on teacher performance on one or more standardized tests including: the Praxis generalist test (Praxis I), Praxis subject tests (Praxis II), the National Teacher Exam (NTE), and in some cases, teachers' SAT and GRE scores. We opted to convert these various test scores into Z-scores in order to place them on a common metric, then took the average Z-score for teachers who had multiple tests listed on their records and experimented with using various test's Z-scores as our measure of teacher academic proficiency.

The process described thus far represents the lion's share of the work that went into constructing a useful dataset for our research. The only remaining tasks were to link our now-linked teacher-student records to NBPTS certification status information from the Educational Testing Service (ETS) , and to link schools to information from the Common Core of Data (CCD) (which includes information on school resources as well as community information from the Census). This was relatively straightforward: we linked teachers by name and schools by National Center for Education Statistics (NCES) code, a coding system that is used for all public schools in the country.

The total process of constructing the North Carolina dataset took approximately one year. During this year a Research Assistant worked nearly full-time on assembling and cleaning the dataset and I spent a considerable amount of time (perhaps as much as 40%) in contact with the state and overseeing the construction of the dataset. Given these figures, the dataset construction likely cost between $100,000 and $200,000 to

[16] This is especially important because measures of teachers' academic proficiency appear to be among the best predictors of teacher quality (Goldhaber, 2002: Ballou and Podgursky, 1997).

[17] This practice of "mean value replacement" is a common statistical technique.

[18] There are 536,736 students who have reading pre and post test scores who have teachers with non-missing teacher test licensure scores and 538,772 students who have math pre and post test scores who have teachers with non-missing teacher test licensure scores. Compared to our NBPTS student samples, 609,160 in reading and 611,517 in math – that is 88.12% of our NBPTS student sample in reading and 88.10% of our NBPTS student sample in math.

construct.[19] This amount of time was necessary because of the difficulties described above, and also because the data we received had state-specific codes, which required significant help from NCDPI for us to understand precisely what particular variables represented.

Thus, we owe a tremendous debt of gratitude to many employees of NCDPI and ETS for answering numerous data questions.[20] And, in particular, the study would not have been possible without the help of Jim Hunt, the former governor of North Carolina. His interest in having an independent evaluation of the impact of the NBPTS program in the state was instrumental to our obtaining the data and receiving timely help from busy state officials.

I believe that North Carolina made a good decision in helping us to engage in this research effort. While it is appropriate for states to be wary about supplying their data to researchers, it is possible to create working relationships that result in new information that states can use to inform policy. States wishing to make policy based on high-quality empirical research therefore should endeavor to maintain high-quality data systems. High-quality student and teacher data is essential for research and accountability purposes. Specifically, to do research on the impacts of teachers on students, it is important to have:

- Annual student testing with achievement tests that are specifically designed to permit value-added modeling;
- Detailed student information including multiple observations of student achievement (preferably annual) and background variables that account for SES;
- Detailed teacher information including degree, experience level, and licensure test performance;
- The ability to link students to their specific teachers (in specific subjects);[21]
- The ability to track student and teacher information;
- And, the ability to link all of the above to other school and community characteristics.[22]

States could provide data as specified above while preserving the confidentiality of students and teachers by developing a data coding system where each student and

[19] Of course these costs depend on the expense of those who are constructing the dataset.
[20] In particular, I'd like to thank Jim Hunt, former governor of North Carolina and founding chair of NBPTS, who encouraged North Carolina public officials to participate in the study; Gordon Millspaugh, formerly of NCDPI, for answering numerous questions about the North Carolina data; and Ashaki Coleman, formerly of ETS, for answering questions about the NBPTS data.

[21] This may mean multiple teachers listed per student in cases where students have different teachers for different subjects different subjects (e.g., in high school).
[22] A more ambitious research agenda would require more comprehensive tracking of teachers. For example, many researchers and policymakers are interested in the value of particular types of teacher training or teacher-training institutions. Research on these issues would require the ability to track teachers from the point when they enter a teacher preparation program all the way to their eventual classroom assignments.

teacher in the public school system is assigned a unique identification number. Such unique ID's could be used for linkage purposes, thus avoiding many of the problems that we faced in our efforts to link by name, and would help to protect confidentiality by allowing all identifying information to be stripped from data being used for research purposes. Some states have already begun this process, and all states should have a vested interest in doing so, as the data they regularly collect for reporting and accountability purposes offer additional opportunities to advance research and ultimately help their educational systems to function more effectively.

References:

Ballou, D. & M. Podgursky (1997). *Teacher pay and teacher quality.* Kalamazoo, Mich.: W.E. Upjohn Institute for Employment Research,.

Clotfelter C.T., H. F. Ladd, and J.L. Vigdor, "Teacher sorting, teacher shopping, and the assessment of teacher effectiveness." Paper presented at the American Association of Public Policy and Management (November 2003).

Fordham Foundation (2004), accessed on December 6, 2004: http://www.edexcellence.net/foundation/global/index.cfm

Goldhaber, D. D. (2002). The mystery of good teaching. *Education Next* *2*(1), 50-55.

Goldhaber, D. D., & Brewer, D. J. (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, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics, 7*(3), 199-208.

Gregorian, V. (2004, November 10,2004). No more silver bullets: Let's fix teacher education*.* *Education Week,* pp. 48, 36.

Hanushek, E. A. A., Rivkin, S. G., & Taylor, L. L. (1996). Aggregation and the estimated effects of school resources. *Review of Economics and Statistics, 78*(4), 611-627.

Hanushek, E. A. (1986). The economics of schooling - Production and efficiency in public-schools. *Journal of Economic Literature, 24*(3), 1141-1178.

National Commission on Teaching and America's Future (2004), accessed on December 6, 2004: http://www.nctaf.org/article/?c=2&body_type=N.

North Carolina School Accountability Reporting (2004), accessed on December 6, 2004: http://www.ncpublicschools.org/accountability/reporting.

Rivkin, Hanushek & John F. Kain, (forthcoming) "Teachers, schools, and academic achievement" ***Econometrica***.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Tennessee Value Added- Assessment System (2004), accessed on December 6, 2004: http://www.state.tn.us/education/tstvaas.htm.