NAEP Background Questions:

What Can We Learn from NAEP About the Effect of Schools and Teachers on Student Achievement?

A Discussion Paper Prepared for the National Assessment Governing Board

September 2002

Michael Podgursky
Department of Economics
University of Missouri
PodgurskyM@Missouri.edu

Note: The author has benefited from comments and suggestions from Michael Ross. The usual disclaimers apply.

"What we cannot speak about we must pass over in silence." -Ludwig Wittgenstein

Introduction

Given the widespread public concern about the performance of public schools, achievement gaps between white and minority students, and the quality of the public school teaching workforce it is not surprising that researchers and policy-makers would seek insight on these matters from the National Assessment of Educational Progress (NAEP). For decades NAEP has served as the national yardstick for charting the progress, or lack thereof, of American k-12 students. NAEP has also called attention to large achievement gaps between black and Hispanic students, and white students. Thus it comes as no surprise that researchers and educational policy-makers would turn to NAEP for answers as to ways to raise student achievement. This motivation – to seek out "contextual factors" associated with student achievement -- has led to the inclusion of over one hundred background variables into NEAP tests and background surveys. The current NAEP includes extended questions asked of students concerning home behavior and their parents, as well as a teacher and a school administrator survey. NAEP "report cards" include a good deal of discussion of these contextual factors, for example, how reading scores vary with student homework and TV viewing, how math scores vary with teacher credentials and with classroom instructional practices. The Department of Education has supported secondary analyses of NAEP by leading educational researchers to help shed light on educational policy issues using these background variables.

In this paper, I briefly review the research literature that has used the NAEP school and teacher background variables to address these educational policy issues.

However, the primary focus of this paper is on the limitations of using cross-section survey data such as NAEP to assess the effect of education policies on student achievement. While NAEP background data may have value as descriptive information concerning student and teacher characteristics and behaviors, one cannot infer causal relationships from simple descriptive cross-tabulations that routinely appear in NCES reports. Nor, in my opinion, is this a problem that can be "fixed" with more sophisticated multivariate statistical procedures such as HLM or structural equation modeling. Convincing research on the effects of education policy variables or school inputs requires true experiments, "quasi-experiments" (naturally occurring approximations to true experiments), or longitudinal student achievement data with good controls for student socioeconomic background, and accurate measures of the relevant policy variables. Unfortunately NEAP falls short in all of these areas.

Inferring Causal Effects of School and Teacher Variables

An early lesson in introductory statistics is that correlation does not imply causality. The point is sometimes illustrated by the story of 19th century Russian peasants who purportedly killed doctors visiting their villages. The peasants noticed that the presence of doctors coincided with the outbreak of disease, interpreted causality in this relationship as running from doctors to disease rather than the other way around, and implemented their own remedy.

Of course we routinely make causal inferences (i.e., X causes Y) based on statistical research. Indeed, if we did not do so there would be little point in conducting statistical research. However, when we make causal inferences usually several

conditions are met. First, the hypotheses tested in the statistical study are built on a body of sound, scientific research. Medical researchers testing the effect of a new drug have reasons for believing that the drug will work. Similarly, educational researchers have sound reasons to believe that that a particular teaching practice will help or harm students. Second, the study uses experimentally generated data. A researcher randomly assigns subjects to one or more treatment groups or a control group. The researcher also controls the "dose" of the treatment variable. Suppose that a researcher is testing the effect of various dosages of a drug to treat colds on a group of patients with colds. A control group of patients receives a placebo. Treatment groups receive different dosages of the experimental drug. Assignment to these groups is random. However, even with random assignment, the researchers will still record the severity of the individual's initial medical condition and collect longitudinal data on the progress of recovery.

If the data from this cold pill experiment are accurately recorded, experimental conditions are maintained, and the sample size is adequate, then the experiment is likely to shed light on the efficacy of the drug. However, even if this experiment is well designed and executed we would want to see the results replicated by other researchers.

Unfortunately, large-scale experiments in education are rare.[1] There seems to be a good deal of reluctance on the part of education policy-makers (and parents) to put children in an experimental situation (although randomized clinical trials are commonplace in medical research -- with much higher stakes). Nonetheless, with the

[1] An exception is the large-scale experiment in class size reduction in Tennessee (STAR) in which several cohorts of elementary school children were randomly assigned to small classes (Finn and Achilles, 1999; Krueger, 1999; Krueger and Whitmore, 2001) However, even here there is some question as to how well experimental conditions were

right non-experimental data a skilled researcher can undertake studies that will support causal inferences concerning education policy. However, such studies of necessity require good data on the socioeconomic background of students, prior student achievement, and need to make sophisticated use of the non-experimental variation in education or school input variables.

Let's return to our example of the cold pill. Suppose that instead of the researcher randomly assigning individuals to treatment and control groups, we rely instead on longitudinal survey data that tracks the course of illness as well as whether these individuals purchased and consumed the cold bills in question. The big difference here is that the patients now self-select themselves into treatment and non-treatment groups (the usual situation in education policy studies). Now suppose we compare the recovery of individuals with colds who buy these cold pills with those who do not. Unfortunately, this simple comparison is likely to lead to incorrect inferences as to the efficacy of the cold pills. For example, it is likely that people with worse colds will tend to buy the medicine, whereas those with milder colds will tend not. This suggests that we need to take account of the severity of the cold prior to the treatment (or non-treatment). Indeed, having the prior medical history of individuals, including the length and severity of prior colds would be very important in producing a rigorous and convincing study.

Even information on initial severity of the cold and prior medical history may not suffice. For example, suppose that poor people are less likely to buy cold pills. That in itself does not bias our findings unless poverty also influences the severity of illness. However, that may very well be the case. It may be the case that on average poor people

---

maintained (Hanushek, 1999). An interesting collection of papers on the use of

take less good care of themselves when they are ill. This in turn prolongs the length of a cold. Thus failing to take account of the education and socioeconomic status of the individuals may be important. It may also be the case that the severity of colds (or the effect of the pills) varies by gender, age, and ethnic or racial group. If we collect prior medical data, and all the other variables that theoretically might affect the length and severity of cold, and we track these individual longitudinally (i.e., over time) we would have a non-experimental database that would help us draw causal inferences about the efficacy of our cold pills.

The U.S. Department of Education and other federal agencies have made major investments in developing these types of detailed longitudinal data files on students. Educational researchers have long made use of large longitudinal databases such as NELS88, High School and Beyond, and Longitudinal Survey of American Youth to analyze the effect of schools and teachers on student achievement. The Department of Education has made a major new investment in the Early Child Longitudinal Survey, which began tracking the educational experience of 23,000 kindergarteners in 1998. These longitudinal databases have been used in a wide range of education policy studies. For example, a large literature has examined the effect of private versus public schools on student achievement using these longitudinal data (Coleman, Hoffer, and Kilgore, 1982; Bryk, Lee, and Holland, 1993; Neal, 1993). Goldhaber and Brewer (1997) examine the effect of teacher characteristics on student achievement using longitudinal student level data (NELS88). Increasingly state-level student assessments will permit development of linked longitudinal student data files. Hanushek, Rivkin, and Kain (1999) make use of a

---

randomized trials in education is found in Mosteller and Boruch (2002).

massive longitudinal database containing several cohorts of elementary students in Texas to examine the effect of variations in teacher quality. Bryk (2002) has tracked the performance of cohorts of Chicago public school students. A similar student data system has been developed in Dallas (Webster and Mendro, 2001). Sanders (2001) has estimated school and teacher effects using longitudinal student achievement data in Tennessee. Angrist and Lavy (2001) use matched school-level dataset of schools in Israel to examine the effect of a teacher training program.

However, even with extensive longitudinal data on households, schools, and students, it still may be the case that there are important omitted variables or factors that bias our estimates of the effect of education policy variables. A promising alternative strategy is to look for "natural experiments." By this we mean looking for cases in which exposure to the "treatment" occurred not through self-selection but through some sort of mechanism that is not systematically related to the outcome we are studying. For example, suppose that we have good longitudinal data on individuals with colds. Suppose that these cold pills in question are not available in some markets. One research strategy then would be to compare individuals who took the pills with observationally equivalent individuals (i.e., severity of cold, demographics, SES) in markets where the pills were not available, or, better still, in the same market before and after the pills became available.

This "natural experiment" approach has been widely used in many economics studies of education and human capital investments. Angrist and Krueger (1991) examined the effect of quarter of birth and differences in state compulsory schooling laws to estimate the effect of an additional year of schooling on earnings. Angrist and Lavy

(1999) examine the effect of arbitrary regulations on class size on student achievement. Hoxby (2000) examines the effect of independent variation in the size of student cohort populations on class size and student achievement. Jacob and Lefgren (2002) use longitudinal student-level achievement data for Chicago public school students and exploit a quirk in the administrative regulation to create a "quasi-experiment" to examine the effect of teacher training on student achievement.[2]

With this general discussion in mind, let's turn to the generic problem of drawing inferences from non-experimental education data. It is useful to write down a simple mathematical equation that represents a "theory" of student achievement. These types of simple models are used in statistical research to estimate relationships and to test hypotheses about causal effects:

Student Achievement in year t = f (Student achievement year t-1, Socioeconomic Factors, Community Factors, School and Teacher Inputs in year t)

This simple model says that student achievement in a given year (t) is determined by student achievement in the prior year (t-1), socioeconomic characteristics of the student, characteristics of the community, and school and teacher inputs in the current year (t). From a policy point of view we are interested in estimating the effect of School and Teacher Inputs (STI's). In principle, STI's are variables that can be changed by educational policy-makers. These might include class size, whether the teacher has participated in professional development, or whether the teacher holds full state certification in her teaching area. With this simple "model" in mind, it is useful to examine the problems with NAEP.

[2] For a more general discussion of these empirical strategies see Angrist and Krueger

1. Poor quality of the socioeconomic data.

In order to accurately estimate the effects of STI's on student achievement, it is important to control for the other factors that influence achievement. One of the most important of these background factors is the socioeconomic status of the students. The socioeconomic background of students has a very powerful effect on student achievement. Statistical studies of student achievement show that of all the variables that affect student achievement, the socioeconomic background of students, particularly variables such as parents' education, have an effect that is consistently large relative to school input variables. This means that if a study of the effect of STI's on student achievement does not control for the socioeconomic background of the student, it is very likely that the estimated effect of the variable in question on student achievement will be biased.

A recent study by Hoxby (2001) highlights the importance of these socioeconomic variables. Hoxby analyzed the effect of family, neighborhood, and school input variables on student achievement and educational attainment using two large nationally representative longitudinal studies of students (the National Educational Longitudinal Survey, NELS88, and the National Longitudinal Survey of Youth, which began in 1979). The list of variables included in each of the areas is extensive. Family variables include parent's education, family income, student race and ethnicity, books at home, etc. STI variables include per-pupil spending, average class size, average teacher

---

(1999).

salary, maximum teacher salary, percent of teachers with MA's, average experience of the teacher, teacher certification status, and other information on school resources. Community variables include income and demographic data on households in the school district and city.

Hoxby compared the percent of the variation in student achievement on various field tests (math, reading) explained by each of these sets of factors. For every test, the percent of the variation explained by the family variables far exceeded the STI variables. The family variables explained from 34 to 105 times as much variation in student achievement test scores as the STI variables. She also examined years of schooling completed at age 33. Family variables explained 19 times as much variation in student educational attainment as did school inputs.

The point here is not that schools or teachers do not matter. Rather it is that the effect of socioeconomic variables is very powerful relative to the measured school and teacher factors. This means that failing to adequately take account of the family background variables in a study can lead to very inaccurate estimates of the effect of STI's if these socioeconomic variables are correlated with STI's.

Consider teacher certification. It is well known that as compared to suburban districts many urban districts have difficulty recruiting certified teachers. It is also the case that children in urban districts are more likely to be poor, black, or Hispanic and that on average poor, black, or Hispanic children tend have lower achievement test scores (whether or not their teachers are certified). Thus, if we simply compare the average achievement scores of students with certified teachers to those of students taught by uncertified teachers, this may reflect the effect of teachers, but it may also reflect the

socioeconomic or racial/ethnic background of the students. Similar problems arise when we compare classroom practices of teachers, a point that we will take up in more detail below.

How good are the NAEP socioeconomic data? My reading of the literature suggests that they are not very good. The socioeconomic data in NAEP come from two sources. First, the students are asked a series of questions concerning their parents and family backgrounds. The most important variables in this regard are parents' education. In principle these would be very useful data for estimating our achievement model. Unfortunately, there seems to be a good deal of measurement error in these student responses. Grissmer, Flanagan, Kawata, and Williamson (1998) have compared the student responses on the family background variables to similar data from the 1990 Census. The NAEP data on race, ethnicity and family type matched the Census data fairly well. However, the information on whether the parents had a college degree was seriously overstated in NAEP. Fourth graders reported 58 percent of parents had college degrees whereas Census data matched to the NAEP sample yielded an estimate of just 25 percent. For eighth graders the divergence was somewhat smaller, but at 42 percent the NAEP value was still well above the Census estimate (25 percent).

Another disturbing factor is the relatively high level of missing values. Grissmer, Flanagan, Kawata, and Williamson (2000, p. 44) report that data on parental education were missing for 36 percent of the fourth graders. If these missing observations are random then estimates of our student achievement model would be unaffected. However, students who do not know or choose not to respond to a question concerning their parents' education are not likely to be a random sample of all students.

A second source of data on students' socioeconomic status comes from the schools. Beginning in 1996, school administrators were asked to provide information on the free and reduced price lunch eligibility for students participating in NAEP. Students are eligible for this program if their family income is less than 150 percent of the poverty line. This is a common measure of student poverty available in state administrative data. However, it is a crude binary measure (eligible/ not eligible) that clearly provides less information than family income or parent's education. It also tends to be underreported at higher grade levels since students at these grade levels do not wish to participate and is also missing for many NAEP students in public schools (Table 1). The rate of missing values for private schools exceeds 50 percent.

2. No Data on Prior Achievement.

In assessing the effect of STI's it is important to take account of prior student achievement. Student achievement in a given year is determined by STI's in the current year, but also by STI's in all of the previous years as well as family inputs in the current and previous years. By including prior test scores in the model, we are capturing the effect of earlier STI's and that permits us to get a more accurate estimate of the effect of current inputs. By controlling for prior achievement, we are assessing the "value added" of the school and teacher inputs provided in the current year.

Consider the effect of teacher classroom behaviors. Suppose we are interested in the effect of a teacher's use of problem sets on mathematics achievement (this is a grade 4 math background question). Our hypothesis is that students whose teachers make greater use of problem sets have greater gains in achievement. In order to test this hypothesis,

we would want to pretest the students in the fall and then retest them in the spring. If the use of problem sets raises student achievement, we would expect to find higher gains in achievement for students whose teachers use more frequent assessments. (This assumes that other confounding factors such as student SES are taken into account.) In this framework, current year teacher inputs are linked to current year student achievement gains.

Now suppose that there are no controls for prior achievement. This permits two sources of bias to taint our estimates of teacher effects. First, students are not randomly distributed across classrooms. It may be that teachers who tend to assign problems sets start the school year with students who are already higher achieving. Perhaps this is because they are in schools or districts with higher-achieving students. When these already above-average students are tested in the spring they will tend to score higher, but we do not know what part of this is due to the fact that they already knew more math in the fall and what part derives from the fact that they learned more during the school year.

A related problem has to do with the timing of the school and teacher inputs. NAEP permits us to correlate current teacher behaviors and current achievement, but it tells us nothing about the behaviors of teachers in earlier years. However, without controls for prior test scores or the school and teacher inputs in earlier years, we cannot be sure whether the higher (or lower) student achievement scores we observe in the spring of 8th grade are due the behaviors of the 8th grade teacher, or those of the 6th and 7th grade teachers.

3. Measurement error on policy variables.

As we will see below, researchers rely on teacher as well as the student responses to measure policy variables. For example, questions of teachers concerning classroom practice are used to asses whether teachers are using certain types of practices favored by reformers (e.g., teaching "higher order thinking," "active learning"). They are also used to assess the effect on student achievement of teacher professional development and teacher education. However, these teacher survey responses may have a good deal of measurement error. A recent study by Mayer (1998) comparing teacher survey responses to actual classroom practices found that responses to specific questions about classroom practices had a good deal of response error. He did find that when individual items were summed along certain broad dimensions, measurement error on individual items was reduced. Ballou (1998) finds considerable measurement error in reports of the teachers' type of certification.

4. Endogenous teacher qualities and classroom behavior.

Another problem in estimating the effect of teachers on student achievement has to do with the direction of causality between teacher behaviors and student achievement. Our model above assumes that the direction of causality runs from teacher classroom behaviors to student achievement. However, one might reasonably argue that educators tailor their teaching practices to fit the academic skills and orientation of their students. Suppose that math teachers with highly motivated students do "hands on" learning and emphasize "higher order" thinking while teachers with weaker students or students prone to misbehavior do more whole class methods and focus on basics. A researcher doing a

cross-sectional study will find a statistical association between teaching practice and student achievement, but causality runs the other direction: student achievement is determining classroom practice and not vice-versa. Once again, the bias caused by this reverse causality might be reduced if measures of students' prior achievement were available (e.g., with longitudinal data). However, such data are not available in NAEP.

Literature Review

With these problems in mind, let's see how well the NAEP literature addresses them.

a. NEAP Report Cards and Related NCES Reports

NAEP Report Cards and related NAEP reports selectively present some of the contextual data in bullets and charts. Two types of data information are presented. The first is the incidence of the particular variable in the teacher or student population. The second is the association between the variable in question and NAEP test scores. For example, the 2000 Reading Report Card included two questions from the fourth grade student surveys concerning how often the teacher helped the student pronounce words and how often the teacher helped them understand new words (every day, once or twice a month, never or hardly ever). We learn that students who report that their teacher helps them pronounce words every day have lower reading scores than students who hardly ever receive such help. Students who say they receive moderate help in learning new words score higher than students who receive a lot of help or those who receive none.

The Math 2000 Report Card presents relatively more data on the credentials and behaviors of classroom teachers. It reports the credentials of the teachers, familiarity with NCTM standards, calculator use, etc. An earlier math report, Hawkings,

Stancavage, and Dossey (1998) reported more extensive cross-tabulations from the teacher survey, including information on teacher majors, certification status, participation in professional development, and teaching experience.

I believe that the information concerning the incidence of instructional practices and credentials of the teacher workforce presented in the Report Cards is useful, although, as noted above, teacher self-reported data concerning detailed instructional practices in the classroom may suffer from a good deal of reporting error. However, I find the discussion in these reports of the relationship between the teaching practices or teacher credentials and student test scores frustrating and rather misleading.

All of these reports contain disclaimers cautioning the reader about making causal inferences from the data and the authors choose their verbs carefully ("… is associated with…, " particular resources "support" student learning). Nonetheless, the reader is left with the impression that that the authors believe that the relationships they are presenting are causal and would like to nudge the reader in that direction as well. What, after all, does it mean to say that a particular STI "supports" student learning? For example, the Math 2000 Report Card and Hawkings, Stancavage, and Dossey (1998) seem predisposed to highlight results suggesting that teachers who know and implement NCTM math standards have higher-achieving students and downplay results that are not consistent with this interpretation.

More generally, the process that seems to drive the selection of variables to be presented and discussed in these reports seems to be whether the result accords with the authors' reading of prior research (broadly-defined) or consensus within the education community. However, it is a simple matter to go through these data and highlight

findings that are inconsistent with received wisdom. For example, in the math report the authors highlight the finding that teachers of 8th grade students who reported "little or no knowledge" of NCTM standards scored lower than teachers who were "knowledgeable or very knowledgeable." However, among 4th grade teachers, students of teachers who were NCTM "knowledgeable" had identical mean scores with those of teachers with "little or no knowledge." NCTM encourages the use of calculators in mathematics instruction, yet the math scores of 8th graders who had access to school-owned calculators was lower than those of students who reported that they did not.

My point here is not to quibble with the NCTM standards or get into the phonics versus whole language debate. Rather it is to point out that the selective interpretation and presentation of the contextual data in NAEP reports is unscientific. The simple cross-tabulations in the NAEP reports tell us nothing about causality. If solid research demonstrates that calculator use raises mathematics achievement scores then that fact stands on its own. The cross-tabulations in NAEP neither reinforce nor undermine that finding. To weave a discussion of research findings into presentation of these NEAP cross-tabulations is misleading and creates the impression that the NAEP data being presented can be used to support or refute hypotheses concerning education policy.

Rather than present these data in such a selective manner, I believe that if NCES continues to collect these contextual data they should simply publish tabulations for all of these contextual data (and their standard errors) in a large set of tables without comment. Alternately, they could simply make them available on the NCES web site (though the "Data Tool" -- a very useful device). The presentation in NCES reports should focus on

overall and subgroup scores. Contextual data should focus only on the incidence of contextual factors but not on the association with student test scores.

b. Analytical Studies

There have been relatively few analytical studies that have made much use of the STI contextual variables in NAEP. In part this may be due to limited access to the student-level NAEP data in the research community due to NCES licensing requirements. However, a more likely explanation would focus on the limitations of cross-section data such as NAEP as compared to richer longitudinal surveys such as High School and Beyond , National Longitudinal Educational Survey, and Longitudinal Survey of American Youth. There have been scores of studies using these surveys (many of which used restricted files) , as compared to a mere handful using NAEP contextual data. Finally, the complex manner in which NAEP student-level plausible values are constructed has clearly deterred some researchers (Barron, 2000; Lee, Croninger, and Smith, 1997).

In these studies, the researchers recognize the importance of controlling for socioeconomic and other background factors in order to get accurate estimates of the effect of school and teacher factors. Several of these studies have aggregated NAEP scores up to units of observation (schools, districts, states) in order to link these units over time to create longitudinal data sets or to match data from other sources. Most have relied on the limited SES data in NAEP in order to control for the effect of student poverty and proceed to examine the effect of contextual variables. For the most part, these researchers have been quite candid about the limitations of cross-section NAEP data for this type of research and have been cautious about making causal interpretations.

Grissmer, et.al. (2001) is a longitudinal study that pools 271 state-level observations of reading and math scores at two grade levels (4^{th} and 8^{th}) between 1990 and 1996. These researchers are interested in identifying which states are making the greatest achievement gains and identifying policies that may be contributing to those gains.[3] In order to do this, they control for the variation in socioeconomic factors between states. These researchers make some use of the NAEP student responses, however they find that the student responses concerning parent's education are highly inflated, thus they rely instead on 1990 Census data. Because they are relying on aggregated NAEP data, they are able to use state-level data on other school resources.

They use a teacher question in NAEP concerning resource adequacy ("I get most of the resources I need." "I get some or none of the resources I need."). These teacher variables from NAEP are occasionally weakly significant, but for the most part are not statistically significant. This may mean that resources do not matter (a position with which the authors do not agree). However, more likely it simply reflects the fact that these NAEP questions are a poor proxy for the variable they really want to measure, i.e., school resources. Random measurement error tends to bias the measured effect of a variable toward zero.

Lee, Croninger, and Smith (1997) examine the effect of "constrained curriculum" within public high schools on 12th grade NAEP math scores for high school graduates using a special version of NAEP that links student records to their school transcripts (the 1990 High School Transcript Study). The High School Transcript Survey sample is much smaller that the full NAEP. This study included only 3,056 high school graduates

[3] Grissmer and Flanagan (2002) update the state performance rankings to include 2000

in 123 schools. By using the transcript data these researchers were at least partially able to overcome the lack of prior student achievement data in NAEP -- their control for prior achievement was the student's 9th grade GPA. If 9th grade GPA were a good proxy for 9th grade achievement, then that would mean the researchers were comparing the mathematics "value-added" of attending particular types of high schools.

The hypothesis examined in this study question is whether the math scores of students are higher if they attended high schools with a restricted curriculum that, in effect, forced all students onto an academic track. This thesis is based on prior work on private schools by Bryk, Lee, and Holland (1993). Using multilevel HLM models they find support for this thesis.

One remarkable feature of this study is the authors' assessment of the quality of the NAEP data they are using. The problem they identify arises from the fact that unlike a traditional survey on student achievement such as High School and Beyond, the achievement scores of all students in NAEP are imputed and not directly measured. This arises from the matrix sampling scheme in NAEP. In order to cover a broad domain of knowledge and hold down the testing time for students, no student is given a complete version of the assessment. In fact, each student takes only a small part of the entire exam. In order to construct a complete student score, NCES uses information contained in the questions the student did answer, along with all of the student background information (e.g., TV viewing, homework, race, ethnicity, gender) to predict a distribution of scores for such a student. They then sample from that distribution and assign those values to the

---

NAEP reading and math assessments.

student ("plausible values," for technical details, see Mislevy, Johnson, and Muraki, 1992).

Lee, Croninger, and Smith believe that the procedure used in the 1990 NAEP to produce "plausible values" creates serious problems for researchers conducting studies such as theirs. They write:

> The National Assessment for Educational Progress represents a major undertaking by both the National Center for Education Statistics in the U.S. Department of Education an the Educational Testing Service, who collect and scale the proficiency scores… They offer solid and current assessments in important subjects across the school curriculum. They include large and nationally representative samples. They are conducted very often. Despite these several advantages, we conclude that there are serious methodological problems that surround the use of NEAP data for the kind of secondary analysis we have attempted here. (Lee, Croninger, and Smith, 1997, p. 115).

It is beyond the scope of this paper to attempt to sort out the implications of using simulated (plausible) values estimates rather than actual observations of student achievement scores for secondary analyses using complicated multivariate statistical procedures that combine conditioned and non-conditioned variables as statistical control variables. Nonetheless, after reading this literature, I believe that it is a topic that deserves further study. "Proceed with caution" should be the order of the day.

Like Grissmer, Raudenbush (2001) seeks to use NAEP to identify states with below and above average school performance. He makes use of the early state-level NAEP data, the 1992 NAEP Trial State Assessment data (TSA). In particular, he focuses on 8th grade math achievement. His work is exploratory, with a goal of illustrating how Hierarchical Linear Models (HLM) can be used to compare the educational performance of states and proposes an innovative use of HLM in this regard. Raudenbush notes that simple comparisons of state NAEP scores to not take account of

significant differences in socioeconomic factors that vary between states. He acknowledges the limitations of TSA for this purpose. "The cross-sectional data of the TSA do not enable the degree of control for prior student achievement that is possible in a longitudinal study such as NELS [National Educational Longitudinal Survey]. Moreover, NAEP indicators of educational policy and practice are not nearly as refined as those in NELS." (Raudenbush, p.6).

Nonetheless, Raudenbush estimates an education production function, treating student achievement as a function of student, school, and classroom factors. Among the classroom factors, he finds that students of teachers who majored in math or mathematics education outperformed other teachers. In addition, teachers who emphasized reasoning or analysis in class had higher-scoring students. Much of the paper is devoted to an exploration of the extent to which parent education affects school resource variables such as school climate and attendance at a school offering 8th grade algebra. The very large size of the NAEP TSA sample (99,980 students in 3,537 schools in 41 states) permits Raudenbush to explore differences between states in equity factors such as the association between parent education or poverty and school climate or algebra classes. He does seem to identify some state-level differences. However, whether these are due to differences in state policies, his preferred interpretation, or simply omitted variables in NAEP, remains to be determined.

Wenglinsky (1997) examines the effect of school resources and spending on student achievement. He aggregates student-level NAEP data up to district-level means and conducts an analysis of 182 school districts. By aggregating to the school district-level he is able to combine NAEP data with district information from Common Core of

Data. He also makes use of the school, teacher, and student surveys from NAEP. In particular, he uses the family background variables from NAEP to control for student SES. The "school environment" variables (which include questions concerning teacher and student absenteeism and tardiness, respect for school property) also capture the SES of the students attending the school. He includes no controls for the race or ethnicity of the student. In the full model there is no evidence that pupil-teacher ratios affect student achievement, however, this may simply reflect the relatively poor quality of the statistical controls for student SES.

The boldest use of the teacher contextual variables is Wenglinsky (2000, 2002). I will devote some attention to these studies since the first of them, published by the Educational Testing Service, was widely publicized and circulated by that organization.[4] Without question it makes the strongest causal and policy claims of any of the studies considered thus far. For example, an ETS press release that preceded the study noted: "…this study shows not only that teachers matter most but how they most matter." How Teaching Matters, p. 32 and ETS Press Release.

The study was released at a media event at the National Press Club and simultaneous press releases by the National Council for the Accreditation of Teacher Education (NCATE) and the American Association of Colleges of Teacher Education (AACTE) claimed that this study yields strong support for the practices of U.S. schools of education.[5]

---

[4] These comments draw on a longer critique (Podgursky, 2001).

[5] "NCATE is pleased to see empirical validation of its standards." Arthur Wise, President, NCATE, NCATE Press Release "Most importantly, the [ETS] study confirms that professional development makes a difference in student achievement." David G. Imig, President, AACTE, AACTE Press Release.

Wenglinsky examines the relationship between the 1996 8th grade science and mathematics test scores of students on NAEP and the characteristics of their 8th grade science and math teachers and their teachers' classroom practices. In the 1996 NAEP science and math teachers were asked about a variety of types of professional development training they received over the last five years. Math teachers were queried on nine types of professional development (e.g., cooperative learning, higher order thinking skills, portfolio assessment, oddly, one of the least common types of professional development for science teachers was laboratories). In addition to professional development, both science and math teachers were asked about 21 different classroom teaching practices. For math teachers, examples included whether they "used a textbook once a week" "addressed routine problems" "assessed students from portfolios at least once a month"

As it turns out, most of the professional development and classroom practices such as cooperative learning and student portfolios had no significant relationship to student achievement after controlling for some student-reported socioeconomic characteristics. After eliminating many statistically insignificant teacher variables, the author found the variables in Table 2 to be significantly associated with student achievement in either science or math. The statistic reported in the table is the estimated effect of a one standard deviation change in the independent variable on standard deviations of student achievement. Thus, other things being equal, emphasis on "higher order thinking" by math teachers raised student achievement by .13 standard deviations, but had no effect on science achievement. Of the six areas of professional development only two had significant positive effects for math teachers, and only one of eight had a

significant positive effect for science teachers. (Interestingly, professional development in classroom management was associated with lower student achievement.) Moreover, a measure of hours of professional development also had no significant effect on student achievement. In general, very few classroom practices were associated with higher student achievement.

What about the notion that "teachers matter most" from the press release? There has emerged a strongly held belief in some parts of the education community that teachers have a larger effect on student achievement than parents or a student's home environment. This belief is fueled by exaggerated claims of the National Commission on Teaching and America's Future in its 1996 report, What Matters Most (i.e., teachers). The quote from the conclusion of this study (featured prominently in the ETS press release) alludes to this stylized fact -- "teachers matter most." However, the statistics in Table 2 simply reinforce what many other studies of student achievement have found, namely, that the socioeconomic background of the student (even when measured with considerable error as in NAEP) has a very large effect on student achievement -- an effect that dominates any other measured school or teacher input.[6]

The author's assertion that he has demonstrated that teachers matter most simply means that a teacher who had the right credentials, professional development, and classroom practice, could, in theory, offset the effect of one standard deviation of socioeconomic disadvantage. But this is a hypothetical exercise. The author has not demonstrated that such high-powered teachers exist in any significant numbers in the

[6] Moreover, we have already seen in Grissmer, et al. (2001) that the student self-reported SES data suffers from serious measurement error. Measurement error means that the effect of SES in Table 2 in underestimated.

population (or, indeed, that any exist in his sample). Nor has he demonstrated that the variation in teacher quality actually observed in the student population has as large an effect an effect on student achievement as variation in SES. That will depend on how the nine variables co-vary with one another. For example, if good practices on one variable are associated with bad practices on another, the net effect on student achievement in the population may be negligible.

In fact, this type of exercise tends to exaggerate the effect of teachers relative to SES effects. The author has sifted through dozens of variables concerning teacher credentials, training, and classroom practice to find nine clusters with the largest effects on student achievement, but he has engaged in no similar search for SES variables. For example, suppose he had information on family income, size, and composition (e.g., female-headed), data from both parents on education credentials and occupation, and information on the SES character of the community in which the student. Now suppose he chose from among the thirty or so SES variables, nine clusters that best predicted student achievement. Without doubt, the SES variables would have a much larger combined effect on student achievement than the teacher variables.

In fact, the author has omitted from his control variables a singularly powerful predictor of student test scores – a student's race. In the 1999 NAEP data, for example, the white-black math test score gap was .74 standard deviations. Studies have demonstrated that SES variables typically explain less than half of white-black test score gaps (Hedges and Nowell, 1998). Thus, had the student's race been included along with the SES controls the total effect size would surely have been much larger.

Not only is the effect of race powerful, but its omission from this study also potentially biases all of the author's estimated teacher effects. For example, are math classes with teachers who emphasize or have had professional development in teaching "higher order thinking skills" disproportionately white? If so, then the positive effects of "higher order" teaching methods or training may reflect nothing more than the racial composition of the classrooms and not the effect of teacher training or practice.

We have already pointed out the limitations of cross-section NAEP data for this type of research. One problem that bears repeating has to do with the direction of causality between teacher behaviors and student achievement. Wenglinsky is assuming that the direction of causality runs from teacher classroom behaviors to student achievement. However, one might reasonably argue that educators tailor their teaching practices to fit the academic skills and orientation of their students. Suppose that math teachers with highly motivated students do "hands on" learning and emphasize "higher order" thinking while teachers with weaker students or students prone to misbehavior tend to do more rote learning and focus on basics. A researcher doing a cross-section study would find a statistical association between "higher order teaching" and student achievement, but causality at least partially runs the other direction: student achievement is causing classroom practice.

Consider how seriously misleading such a cross-section approach would be in a medical study. Doctors choose their treatments based on the condition of a patient. Suppose that doctors treating breast cancer chose localized surgery for patients with small localized tumors but radical mastectomies for patients with a more advanced and life threatening forms of the disease. If we simply regress patient survival rates on

treatments, with no control for the pre-existing condition of the patient, we would erroneously conclude that radical mastectomies reduce the survival probabilities of a patient. By analogy, the bias in a student achievement study caused by the fact that teacher tailor instructional strategies to students can be mitigated if measures of students' pre-existing achievement are available (e.g., with longitudinal data), however such data are not available in a cross-section survey such as NAEP.

These are the major studies of which I am aware that attempt to estimate the effects on student achievement of the NAEP school and teacher contextual variables. If we cast the net a bit wider, another study deserves mention. Swanson and Stevenson (2002) make extensive use of the teacher classroom behavioral variables to assess whether state attempts at education reform have trickled down to the classroom. They make no attempt to assess the effects of either state policy activism or NAEP teacher variables on student achievement. Interestingly, they conduct their study by linking schools that had participated in both the 1992 and 1996 state NAEP math assessments. Not surprisingly, the lion's share of variation in teacher instructional practice is within schools or between schools within a state. Only three percent comes from between state variation. Nonetheless, they do find evidence of a positive effect of state-level activism.

Given this the extensive educational data collected by states and in other Department of Education files such as Common Core of Data and Schools and Staffing Surveys, one promising road to increasing the analytical value of NAEP is to develop methods to link these files. An innovative study in this regard is McLaughlin and Drori (2001), who link school-level student achievement data from twenty state assessments to the very extensive data on schools and teachers in the Schools and Staffing Surveys.

They use NAEP to develop a common interstate achievement metric. In principle these methods could be used to link student or school level longitudinal files across states as well.

Conclusion

To answer the question posed in the title of this essay – what can we learn from NAEP about the effect of schools and teachers on student achievement? Unfortunately I must conclude – not a great deal.

Returning to our medical analogy, I think that the primary value of the NAEP survey is to monitor the academic health of American children. Some contextual data is important in this monitoring function. We must be able to accurately measure the performance of population subgroups by gender, race, ethnicity, poverty, rurality, etc. Medical surveys are invaluable in identifying problems and highlighting where resources should be targeted. However, they cannot tell us what will cure an illness. For that we need to turn to different types of data. In medicine these are primarily experimental data generated from controlled clinical trials. In education, the data collected are primarily non-experimental, but can be used in ways that approximate experimental conditions in some respects. However, asking NAEP to serve both functions risks dilution of its unique and vital monitoring role. A fixed NAEP budget imposes tradeoffs. The opportunity cost of a parent survey or more teacher contextual data is a smaller sample size. Alternately, fewer teacher contextual questions would permit an increase in the size of the student sample, and hence more accurate measures of subgroup achievement scores nationally and state-by-state.

As long as these teacher and school contextual data are collected as part of NAEP there is going to be a strong temptation to draw inferences about their link to student achievement. In this paper I have argued this is simply not possible given the inherent

limitations of a cross-section survey like NAEP. NCES and NAGB should resist the temptation, follow the counsel of Wittgenstein, and remain silent on the matter.

The research center of gravity for quantitative education policy research is clearly with national longitudinal surveys, and increasingly, linked student or school-level longitudinal data based on district or state assessments. This is an efficient allocation of research resources. Nearly all states now conduct state assessments of student achievement. As a result, a vast amount of data is becoming available from states concerning school resources, teachers, and student achievement. The No Child Left Behind Act will undoubtedly accelerate the collection and analysis of these data. NAEP national and state data will remain an indispensable monitor of state and national performance, and will permit comparisons of achievement gains across states. However, quantitative policy research on the effect of teacher and school inputs on student achievement will not rely on NAEP.

Table 1

Percent of Student Records
Missing Lunch Program Eligibility:
Public Schools
2000 National NAEP

	Percent Missing Lunch Program Eligibility
Grade 4 Math	13 %
Grade 8 Math	16 %
Grade 12 Math	24 %
Grade 4 Reading	9 %
Grade 8 Reading	12 %
Grade 12 Reading	13 %

Source: NAEP Data Table. http//nces.ed.gov/nationsreportcard/neapdata/

Table 2

Standardized Effects of Teacher Inputs, Professional Development, and Classroom Practices on Mathematics Achievement: Relative Scores

Factor	Math	Science
Student socioeconomic status	.76	.75
Average Class Size	.10	.11
Major/minor in Subject Area	.09	.09
Professional Development in working with different student populations	.21	.00
Professional Development in higher-order thinking skills	.12	.00
Professional Development in laboratory skills	n/a	.13
Professional Development in Classroom Management	.00	-.13
Classroom practice variables: Hands-on Learning	.25	.18
Higher-order thinking skills	.13	.00
Assessment without testing	-.18	.00
Assessment with testing	.00	.21

Source: Reproduced from Wenglinsky (2000), Table 3

References

Angrist, Joshua and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." in Orley Ashenfelter and David Card (eds.) Handbook of Labor Economics Vol. 3. Netherlands: Esevier Science, pp. 1277- 1306.

Angrist, J., and A. Krueger. 1991. "Does Compulsory Schooling Affect Schooling and Earnings?" The Quarterly Journal of Economics. Vol. 106, No. 4, November, 979-1014.

Angrist, Joshua and Victor Lavy. 1999. "Using Maimonnides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." Quarterly Journal of Economics Vol. 114 No. 2 (May), pp. 533-575.

Angrist, Joshua and Victor Lavy. 2001. "Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools." Journal of Labor Economics. Vol. 19 No. 2 (April), pp. 343-69.

Ballou, Dale. 1998. "Alternative Certification: A Comment." Educational Evaluation and Policy Analysis. Vol. 20. No. 2 (Winter), pp. 313-315.

Barron, Sheila. 2000. "Difficulties Associated with Secondary Analysis of NAEP Data." in Nambury S. Raju, James W. Pelligrino, Meryl W. Bertenthal, Karen J. Mitchell, and Lee R. Jones (eds.) Grading the Nation's Report Card: Research Evaluations from NAEP. Washington, DC: National Academy Press. pp. 172-195.

Bryk, Anthony S. 2002. "No Child Left Behind Chicago Style: What Has Really Been Accomplished?" Harvard University. Program on Education Policy and Governance. PEPG/02-05.

Bryk, Anthony S., Valerie E. Lee, and Paul Holland. 1993. Catholic Schools and the Common Good. Cambridge, MA: Harvard University Press.

Coleman, James S., Thomas Hoffer, Sally B. Kilgore. 1982. High School Achievement: Public, Catholic, and Private Schools Compared. New York: Basic Books.

Finn, Jeremy D. and Charles M. Achilles. 1999. "Tennessee Class Size Study: Findings, Implications and Misconceptions." Educational Evaluation and Policy Analysis. Vol. 20. No. 2 (Summer),

Goldhaber, Daniel and Dominic J. Brewer. 1997. "Why Don't Schools and Teachers Seem to Matter?" Journal of Human Resources Vol. 32 No. 3 (Summer), pp. 505-523.

Grissmer, David, Ann Flanagan, Jennifer Kawata, Stephanie Williamson. 2000. Improving Student Achievement: What State NAEP Test Scores Tell Us. Santa Monica, CA: Rand.

Grissmer, David and Ann Flanagan. 2002. "Tracking the Improvements in State Achievement Using NAEP Data." Presented at Research Seminar II. Instructional and Performance Consequences of High-poverty Schooling. Washington DC: March 11.

Hanushek, Eric. 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. Vol. 21 No. 2 (Summer), pp. 143-163.

Hanushek, Eric A, John F. Kain, Steven G. Rivkin, 1998. "Teachers, Schools, and Academic Achievement." NBER Working Paper 6691. Cambridge, MA: National Bureau of Economic Research.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." in Orley Ashenfelter and David Card (1999) Handbook of Labor Economics. Netherlands: Elsevier Science.

Hedges, Larry V. and Amy Nowell. 1998. "Black-White Test Score Convergence since 1965." in Christopher Jencks and Meredith Phillips (eds.) The Black-White Test Score Gap. Washington DC: Brookings Institution.

Hoxby, Caroline. 2001. "If Families Matter Most, Where Do Schools Come In?" in Terry Moe (ed.) A Primer on America's Schools. Stanford University: Hoover Institution Press, pp. 89-126.

Hoxby, Caroline. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." Quarterly Journal of Economics. Vol. 115 No. 4., pp. 1239-1286.

Jacob, Brian A. and Lars Lefgren. 2002. "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago." NBER Working Paper 8916. Cambridge, MA: National Bureau of Economic Research.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." Quarterly Journal of Economics Vol. 114 No. 2 (May). pp. 497-532.

Kreuger, Alan B. and Diane M. Whitmore. 2001. "The Effect of Attending a Small Cass in the Early Grades on College Test-Taking and Middle School Test Results: Evidence from Project STAR." Economic Journal. Vol. 111 (January), pp. 1-28.

Lee, Valerie, Robert G. Croninger, Julia B. Smith. 1997. "Course-taking, Equity, and Mathematics Learning: Testing the Constrained Curriculum Hypothesis in U.S. Secondary Schools." ?" Educational Evaluation and Policy Analysis. Vol. 19 No. 2 (Summer), pp. 99-121.

Mayer, Daniel P. 1999. "Measuring Instructional Practice: Can Policymakers Trust Survey Data?" Educational Evaluation and Policy Analysis. Vol. 21 No. 1 (Spring), pp. 29-45.

Mislevy, Robert J., Eugene G. Johnson, and Eiji Muraki. 1992. "Scaling Procedures in NAEP." Journal of Education Statistics. Vol. 17 No. 2 (Summer), pp. 131-154.

Mosteller, Frederick and Robert Boruch (eds.) 2002. Randomized Trials in Education Research. Washington DC: Brookings.

Neal, Derek. 1997. "The Effect of Catholic Secondary Schooling on Educational Attainment." Journal of Labor Economics. Vol 15, pp. 98-123.

Podgursky, Michael. 2001. "Flunking ETS: A Review of How Teaching Matters." Education Matters. Vol. 1 No. 2 (Summer), pp. 75-78.

Raudenbush, Steven W. "Synthesizing Results from the NAEP Trial State Assessment." In David W. Grissmer and J. Michael Ross (eds.) Analytic Issues in the Assessment of Student Achievement. U.S. Department of Education: Washington, DC. , pp. 3-42,

Raudenbush, Steven W., S. W. Fotiu, Y. F. Cheong "Synthesizing Results from the Trial State Assessment." Journal of Educational and Behavioral Statistics.

Raudenbush, Steven W., S. W. Fotiu, Y. F. Cheong "Synthesizing Results from the Trial State Assessment." Educational Evaluation and Policy Analysis.

Sanders, William L, Arnold M. Saxton, and Sandra P. Horn. 1997. " The Tennessee Value-Added Assessment System: A Quantitative, Outcome-Based Approach to Education Assessment." in Jason Millman (ed.) Grading Teachers, Grading Schools. Thousand Oaks, CA: Corwin Press, pp. 137-162.

Swanson, Christopher and David Lee Stevenson. 2002. "Standards-Based Reform in Practice: Evidence on State Policy and Classroom Instruction from the NAEP State Assessments." Educational Evaluation and Policy Analysis. Vol. 24 No. 1 (Spring), pp. 1-27.

U.S. Department of Education. National Center for Education Statistics. 2000. School-level Correlates of Academic Achievement: Student Assessment Scores in SASS Public Schools. NCES 2000-303. by Donald McLaughlin and Gill Drori. Project Officer, Michael Ross. Washington, DC.

Webster, William J. and Robert Mendro. 1997. "The Dallas Value-Added Accountability System." in Jason Millman (ed.) Grading Teachers, Grading Schools. Thousand Oaks, CA: Corwin Press, pp. 81-99.

Wenglinsky, Harold. 1997. "How Money Matters: The Effect of School District Spending on Academic Achievement." Sociology of Education Vol. 70. No. 3 (July), pp. 221-237.

_____. 2000. How Teaching Matters: Bringing the Classroom Back Into Discussions of Teacher Quality. Princeton, NJ: Educational Testing Service.

_____. 2002. "How Schools Matter: The Link Between Teacher Classroom Practices and Student Academic Performance." Education Policy Analysis Archives. Vol. 10. No. 12 (Feb. 13). http://epaa.asu.edu/epaa/v10n12.