Estimating Relationships in NAEP: A Comparison of IV and Traditional Methods

Duncan Chaplin
The Urban Institute

September 23rd, 2004

Abstract

As researchers are keenly aware, controlling for statistical bias is critically important, yet often difficult to do effectively. Without proper controls we could easily come to faulty conclusions about the directions and magnitudes of estimated effects. When analyzing NAEP data, researchers have used a variety of methods to control for bias. These include multivariate regression to control for observable student, school, district, and state characteristics, as well as analyses that investigate variation over time, across cohorts, and within students. One method that has seldom, if ever, been used with NAEP data is the Instrumental Variable method (IV), which is playing an increasingly prominent role in the field of education policy research, especially among economists. This study investigates the feasibility of using IV for estimating impacts of a variety of education policy variables.

A subset of feasible models are chosen for estimation and IV estimates are compared to estimates produced using a number of other methods. When comparing and evaluating the feasibility of the alternative methods I consider the following issues: bias in slope estimates caused by omitted variables, bias caused by measurement error in the education policy variables, precision of the slope estimates, variation in treatment effects, and data availability. The resulting work should expand the ways in which NAEP data can be used and the ways in which impacts of education policy variables can be estimated.

Background

The term "scientifically based research" appears over 100 times in the new No Child Left Behind Act (Olson and Viadero, 2002).[1] While the exact meaning of this phrase is far from clear, what is certain is that the pressure to produce more rigorous research to guide policy has been ratcheted up. Much of this work will require the implementation of new studies, many of which will be based on true experiments. Indeed, the U.S. Department of Education is greatly increasing the amount of money it will be spending on experiments (Viadero, 2002). However, such research will take time and will likely be quite expensive. In addition, because of the expense, it is likely that such research will only be able to address a limited set of policy research questions (Manski and Garfinkel, 1991; Heckman and Hotz, 1989). Consequently, it is becoming increasingly important to search for new and innovative ways to use existing data.

The National Assessment of Educational Progress (NAEP) was instituted in 1969, over a third of a century ago, to help measure the nation's academic progress. While measuring progress remains its primary function, secondary analyses of NAEP data are also considered important (Raju et al., 2000). Indeed, many researchers who were funded for past NAEP Secondary Analysis Grant projects proposed to look at relationships between policy-related variables and student test score outcomes. While few researchers would claim to have strong evidence of causality based on such research, most attach at least some weight to the possibility that their results suggest the policy variables they analyze have real impacts. Most of these projects, and research on NAEP in general, looks at cross-sectional associations or, in some

[1] The 2001 reauthorization of the Elementary and Secondary Education Act that covers all federal funding of education in the U.S.

cases, changes over time in relationships between various variables.[2] None of the studies to date, however, uses Instrumental Variables.[3] For these reasons, developing improved methods of estimating relationships between NAEP test score outcomes and various education policy variables that are free from bias caused by unobserved differences between different sub-groups of students could be extremely beneficial.

A variety of methods are used to control for bias in education policy research. Instrumental Variables (IV) methods are playing an increasingly prominent role in this area, especially among economists. This method is sometimes described using the words "quasi-experimental" or "natural experiment" (Angrist and Krueger, 2001). Thus, it seems likely that, at least in some cases, research using the IV method might help satisfy the goal of the No Child Left Behind act (NCLB) for more rigorous research.

This paper investigates various ways in which NAEP has and could be used to estimate relationships between various education policy variables and student test scores, with a particular focus on the methodologies currently used most commonly with NAEP data, and some more innovative methods, such as those based on Instrumental Variables. The strengths and

---

[2] Examples of relevant projects include those of Desimone, Braun, and Curry funded in 2002; Grissmer in 2001, 2000, and 1999; Von Secker in 2001; Huang and Sloop in 2000; Guthrie in 1999; Niemi in 1997; Wise in 1996; Wong, Franks, Williams, Davidson & Davenport, and Lee in 1995; and Wainer in 1994.

[3] Many economists who use instrumental variable methods to estimate relationships in education policy research release their reports as National Bureau of Economic Research working papers before formal publication. As of February 2003, only one paper on this website used NAEP test score data as the major outcome (Fuchs & Reklis, 1994).

weaknesses of each method are compared based on estimating effects of a small set of education policy variables, and recommendations are made on how to proceed in the future.

A number of methods are considered:

1) Cross-sectional,
2) Differences over time,
3) Differences across cohorts,
4) Differences within Students,
5) Differences within Schools, and
6) Instrumental Variables.

In comparing these methods I consider a variety of issues. These include the following:

1) Potential bias in slope estimates caused by unobserved factors that affect the outcomes and are correlated with the education policy variables of interest,
2) Potential bias in slope estimates caused by measurement error in the policy variables,
3) Precision of the slope estimates,
4) Variation in treatment effects, and
5) Data availability.

I begin with a discussion of the types of models that are considered in this analysis, giving particular attention to analyses that have been done with NAEP data and a number of potential examples of ways in which the Instrumental Variable method could be used. I then discuss the types of issues that should be considered when comparing these models and how they are dealt with in this study.

Analysis Methods

When estimating effects on student achievement, it is important to control for the fact that student backgrounds differ greatly and are likely to vary with many of the education policy variables being analyzed. Most previous research in this area has addressed this issue by controlling for observed measures of student background factors (e.g. parent education, race, and gender), and, when possible, previous academic achievement (test scores). NAEP data include information on some of these background factors but lack information on prior achievement at the student level. This has limited the usefulness of NAEP data for estimating causal effects. In this paper I use methods that both allow us to estimate effects using NAEP data and produce alternatives to estimates based on the standard "value-added" methods, that control for previous test scores.[4]

Cross Sectional

A large number of researchers have conducted cross-sectional analyses of NAEP data to analyze relationships between various education policy variables and student test scores, or closely related outcome measures. (e.g. Huang and Yu, 2002; Sloop, 2002; Guthrie et al., 2001; Wenglinsky, 1998;and Fucks and Reklis, 1994) and others have proposed such estimation (Wong in 1995). In order to make strong causal statements based on such models,[5] one must

---

[4] Meyer (1996) gives a good discussion of value-added methods.

[5] Many researchers, especially in the field of education research, use Hierarchical Linear Models (HLM). These models help adjust standard errors of slope estimates for potential bias caused by correlations between observations and produce more efficient slope estimates, but do not adjust for the bias in slope estimates being discussed here. In addition, HLM models can be written as single equations, similar to those being presented here and the same points apply.

assume that the error terms are uncorrelated with the education policy variables being included in the analysis. To be more precise, suppose we have the following model:

$$Y = E'\beta 1 + X'\beta 2 + \varepsilon$$

where Y = the outcome in question, typically a test score,

E = the education policy variable,

X = the control variables[6],

ε = unobserved factors that also impact the outcome, and

$\beta 1$ and $\beta 2$ are parameters to be estimated.

Researchers have used this model to estimate impacts of a number of education policy variables with NAEP data. These policy variables include those designed to equalize spending across districts (Wenglinky, 1998), district current expenditures per pupil and discretionary rates for instructional expenditure (Huang and Yu, 2002),[7] various teacher and school characteristics (Sloop, 2002), opportunities to read and engagement in reading (Guthrie et al., 2001), and additional school characteristics (Fuchs and Reklis, 1994). In each case, the researchers are implicitly assuming that cov $(\varepsilon, E|X)=0$[8] when they use their results as evidence of causal relationships. Such statements are not uncommon in the literature. For instance, based on their cross-sectional results, Guthrie et al. (2001) conclude that "Classrooms and schools should invest

[6] In many analyses using NAEP data these would include a student's race, gender, and parent education levels.

[7] Interestingly, Huang and Yu used data from 1990, 1992, and 1996, meaning that they could have run all three years together and controlled for state dummy variables which would have had the effect of controlling for any factors that remained fixed within states over time.

[8] This means that the covariance between ε and E equals 0 conditional on X.

time and resources toward increasing reading engagement," and that "Teachers who afford students opportunity to read build a context for engaged reading, which increases achievement."

Whether or not the assumption above can be justified is a difficult question to answer. Ideally, one would estimate impacts of education policy variables using true experiments, as the Department of Education is hoping to do in a number of cases, and then compare the results to those using standard methods. However, this can only be done for a limited number of topics as experiments are quite expensive. Unfortunately, a great deal of evidence suggests that when experiments can be done, they often do not yield the same results as non-experimental analyses (Bloom et al., 2002; Agodini & Dynarski, 2001; and Wilde & Hollister, 2002). Fortunately, even without experimental data, there are non-experimental methods that can be used to estimate effects under less restrictive assumptions than the standard assumptions, and to improve upon the resulting estimates, under certain conditions.

One example of this is when one is interested primarily in an interaction effect, rather than in the main effect of a policy variable. For instance, Wenglinsky (1998) focuses much of his discussion on the fact that districts with higher expenditures at the school level have smaller test score gaps by socio-economic status (SES) of the students.[9] Given this focus, he could have made a slightly weaker assumption than the one given above. Let,

$$Y = E' \beta 1 + X'\beta 2 + S'\beta 3 + S*E'\beta 4 + \varepsilon$$

where S = SES and $\beta 4$ is the parameter of interest.

Now suppose that $cov(\varepsilon, E|X) <> 0$. This means that the estimate of $\beta 1$ will be biased. However, it may still be possible to estimate the interaction between the policy variable (E) and

[9] SES is a composite measure of the social and economic well-being of a family, typically based on education, income and occupation.

student SES (S) as long as the degree of bias in the estimated effect of E does not vary with SES. More precisely, it would be sufficient to assume that cov(S*E, ε|E,Z,and S)=0. Suppose that SES were an indicator variable designating a student as either high or low SES. The estimated impact of expenditures may be biased for both high- and low-SES students, but as long as the magnitude of the bias is the same, we can still get a true estimate of the *differential* impact by comparing the estimated impacts for the high and low-SES students, i.e. by estimating the interaction term.

Differences over Time

Looking at differences in impacts between different subsets of data is, indeed, a fairly common way of getting rid of bias. However, rather than compare high- and low-SES students at a given point in time, many researchers prefer to get rid of bias by looking at differences over time. For instance, Chaplin et al. (2003) use school enrollment data by state and year, to investigate impacts of the minimum wage. A number of researchers have also looked at variation over time in NAEP. State-level data over time has been used by Amrein & Berliner (2002) to estimate impacts of high-stakes testing policies and Yang (2002) for impacts of Statewide Systemic Initiatives (SSI). One study (Swanson and Stevenson, 2002) has even analyzed changes over time at the school level to estimate impacts of various Standards Based Reform policies in NAEP.[10]

[10] This is possible for a subset of schools that happen to be in subsequent years of the data.

The beauty of these methods is that they can implicitly control for all factors that remain fixed within a geographic region and at a specific time by including geographic and year dummy variables in the analyses.[11] In other words, suppose that,

$$Ygt = Egt' \beta 1 + Xgt'\beta 2 + \alpha g + \gamma t + \varepsilon gt$$

where g represents the geographic region (state or school in the examples here),

t represents time,

αg = factors that affect outcomes in region g (and do not change over time), and

γt = factors that affect outcomes at time t (and do not vary across regions).

Now, if we estimate this model using cross-sectional data, the resulting estimates might be biased if αg and/or γt were correlated with Egt. However, by using data that varies both over time and across regions, we can control for both αg and γt using dummy variables and thus get rid of these potential sources of bias.

Another interesting use of NAEP data over time is by Cook and Evans (2000) who look at test score differences by race, time, and school.[12] They find that most of the change over time in the black/white test score gap is explained by a drop in test score differences between blacks and whites attending the same schools, rather than by a drop in test score differences between schools. Because they focus on racial differences, they need not be able to estimate the impacts of attending a given school on Blacks or Whites, as long as the bias is the same for both groups. In addition, because they are looking at changes over time, this bias could be different for Blacks

[11] Many of these researchers did not explicitly use such controls, though in some cases their methods implicitly controlled for many of the same types of bias.

[12] Cook and Evans (2002) do not follow individual schools over time. Rather they look at the fraction of variance explained within and between schools at two points in time.

and Whites, as long as it doesn't change over time differently for Blacks and Whites.[13] Thus, by looking at a triple difference (race, school, and time) they can get rid of even more sources of bias than the standard change over time analyses.

Differences across Cohorts

A number of researchers have also used NAEP to follow cohorts of students over time. This is possible because NAEP covers students in 4th and 8th grades and if exams in the same subject are given 4 years apart then they are, more or less, randomly sampling from the same cohort of students. Based on this type of analysis Amrein & Berliner (2002) find some evidence of improving reading scores for those states implementing high-stakes tests. This method gets rid of bias in a manner similar to the change over time method, but instead of controlling for all time specific sources of bias (γt above), it controls for all cohort specific factors.

Differences within Students

A particularly interesting variation on the differences methods discussed above was proposed by Wise and Abedi in their 1996 NAEP secondary analysis grant abstracts. They proposed to look at variation within students depending on when a given item is given on the

[13] If, however, selection of who attends which school changes differentially by race over time, then their results may be incorrect. Suppose, for instance, that in the earlier time period, children of low-skilled blacks attended schools with relatively well-off whites resulting in large within school gaps whereas by the end of the decade income segregation replaced racial segregation so that only children of higher-income blacks could attend school with well-off whites. If this were the case then the black-white gap within schools might have dropped not because of any real changes in the impacts of these schools, but rather because of a reduction in within school skill differences that existed prior to entering school.

exam. By focusing their analyses on within student variation on different items of an exam, they can, in effect, control away for any factors that would impact that students performance equally on all items. Let,

$$Yij = Eij' \beta 1 + Xij'\beta 2 + \alpha i + \gamma j + \varepsilon ij$$

where i stands for individual i, j for subject j, αi represents all factors that affect individual i's general learning skills and γj represents all factors that affect performance in subject j equally for all students. The αi could include impacts of, for instance, almost all parent and student characteristics, including many that would not be available in NAEP (parent academic skills) and many that would probably not be available in almost any dataset (parent wealth). Thus, by being able to control away for αi, and γj this method represents a particularly powerful method of controlling for unobserved factors.

This particular method suggests an important possibility for NAEP—that it could be used to analyze impacts of various policies on particular types of questions, controlling for an individual student's overall level of performance. Of course, its usefulness will be limited by the fact that it can only be used to estimate impacts of policies that can be clearly argued to impact one type of learning more than another, and it can only be used to estimate this differential impact.

Differences within Schools

Cross-sectional analyses of NAEP data using policy variables measured at the student (or classroom) level are, in effect, using both the within and between school variation to estimate policy impacts. For instance, Wenglinsky (2002) uses NAEP data to estimate impacts of teacher characteristics and practices on NAEP test scores. Because he has multiple teachers in many schools he could have controlled for school dummy variables and thereby got rid of any potential

bias caused by all school-level factors that affect all students within a school equally. However, this might result in even more biased estimates if, for instance, there were substantial tracking within a school.[14] Fortunately, the NAEP teacher and school questionnaires include information about tracking. This information can be used to limit the sample to untracked courses. Any remaining variation in teacher characteristics may be more or less random and, therefore, produce unbiased estimates of the effects of teacher characteristics.[15] This method is described in more detail in Appendix B.

Instrumental Variables

While the methods based on differences described above can be very useful for controlling for statistical bias, there are many cases where researchers suspect that omitted variables may vary in the same ways with the policy variables (across time, cohort, sub-group, etc.), meaning that biased estimates are still possible even after using the more sophisticated methods discussed above. For instance, Amrein and Berliner (2002) note that some states (e.g. North Carolina) aligned their state tests with NAEP before others so that estimates of state

---

[14] The odd result of increased bias after controlling for school dummy variables can result because some of the variation in teacher skills between schools might be uncorrelated with prior student achievement and, therefore, be helpful for estimating an unbiased effect of teacher characteristics on student outcomes. By including school dummy variables we would get rid of this helpful variation as well as any school-level variation that might be causing bias.

[15] In addition, one could test to see if the remaining variation were more or less random by estimating the associations of teacher characteristics with student race and gender, as discussed in Appendix B.

policies based on changes over time could miss the impacts of these policies if the analysis does not account for the timing of alignment. Similarly, change over time results can also be biased by the fact that Special Education and LEP students are often excluded from taking NAEP. Indeed, these rates have varied over time differently by state in ways that may explain differential observed changes in performance levels of some states, especially North Carolina and Texas during the 1990s (Amrein & Berliner, 2002 and Haney, 2001).

For these reasons alternative methods may still be helpful. One particular alternative, often used by economists,[16] is to use Instrumental Variables to estimate the impact of a policy variable on some outcome of interest. In order to justify this method one need not assume that the policy variable is uncorrelated with the omitted variables. Rather, one need only have some instrumental variable, say Z, that has no direct impact on the outcome and is not correlated with any omitted variables, but is reasonably well correlated with the policy variable. We do not need to be able to estimate an unbiased impact of the instrumental variable on the education policy variable—we only need to know that the two are strongly related to each other.[17] Given such a variable, one can estimate a two-stage model where the first stage estimates impacts of Z on E and the second stage estimates the impact of E on Y (using a predicted value for E).[18]

Following are some examples of various policy relevant variables that might be possible to analyze using NAEP data and Instrumental Variables, with a brief description of how each study might be conducted.

---

[16] See Goldberger (1991) for a good discussion of this method.

[17] The statements in this paragraph are conditional on X.

[18] In practice these two equations are generally estimated jointly.

School Choice Policies

Hoxby (2000b) argues that having more districts in an area can increase the competition between these districts and, consequently, performance levels. She estimates a number of models, including one which uses the number of rivers and streams in the area as the instrumental variable for the number of school districts (more rivers and streams are found to be associated with more school districts). She reports positive impacts on student performance in the National Educational Longitudinal Survey in grades 8, 10, and 12 for non-minority students. Similar models could be estimated using NAEP data, but covering a far larger number of school districts and, consequently, have the possibility of yielding more precise estimates.

School Type

Neal (1997) estimates the impacts of attending a Catholic school, as opposed to a regular public school, on student performance as measured in the National Longitudinal Survey of Youth. His instrumental variables for attending a Catholic school are the population density of Catholics in the locality and Catholic schools per square mile. He finds positive impacts on graduation rates but no impacts on test performance. Using NAEP would allow for a much larger sample of schools and the addition of more years of data.

Peer Effects

A great deal of the education research essentially ignores the potential importance of peer effects. For instance, much of the literature suggesting that private schools perform better than public ones does little (if anything) to control for peer effects. Unfortunately, controlling for these effects is extremely difficult (Manski, 1993). Hoxby (2000c) and Hanushek et al. (2001) estimate the importance of peer effects on individual performance controlling for fixed effects for each individual student, school, and school by grade. They are able to do this because they

have panel data on individual students over time. Since NAEP does not follow students over time it would not be possible to replicate their methods. However, there is a related type of estimation that would be possible using NAEP data and is very similar to the method used by Hoxby (2000c). In particular, one could use variation in race and gender composition over time as an instrumental variable for peer effects, as measured by the average performance levels of current students.[19] In addition, one could follow a subset of individual schools over time (similar to Swanson and Stevenson, 2002) and thereby control for school fixed effects while doing the IV method.

Unionization

Hoxby (1996) also used changes in laws regulating whether or not unions are allowed in a given state as instrumental variables for unionization and found negative impacts—i.e. more unionized areas tended to have higher dropout rates. Similar methods could be used to analyze impacts of unions on student achievement in NAEP.

Teacher Qualifications

It is also possible to estimate impacts of teacher qualifications on student achievement using an IV approach. States have various qualification requirements for teacher licensure.[20] They also have various reward structures for teachers obtaining these credentials and

---

[19] I would control for past values of the average racial and gender balance at the school in both stages of the estimation as well as the race and gender of each student in the data.

[20] Data on some of these requirements are available in Quality Counts (1998, 1989, 1999, 2000) and Golhaber and Brewer (1999). I would also attempt to collect more information on these rules and regulations for the proposed project.

experience.[21] I can, therefore, use the variation in these credentials and rewards as an instrumental variable for teacher credentials in a two-stage model. More precisely, one can obtain data on these factors for multiple years and then estimate the models using the changes in these factors over time, controlling for both state and year dummy variables in both stages of the analysis.

Class Size

Angrist and Lavy (1999), Angrist & Lang (2002), and Hoxby (2000a) use institutional controls on class size to generate instrumental variables that can be used to estimate the impacts of class-size on student outcomes.[22] They are able to do so largely because they look only at school systems within which the rules are constant (in Israel and in Brookline, MA). This would be a non-trivial exercise in the U.S. as there are over 16,000 school districts, each of which could

[21] Salary schedules, typically based on experience, vary by school district. Lacking district-level information, I might rely on the average returns to credentials and experience at the state level, as reported in the NCES Schools and Staffing Survey.

[22] In particular, they note that as the number of students in a given grade at a given school increases, rules governing class size change the number of students in a classroom in very non-linear ways. For instance, if the maximum class size is 25 then moving from 24 to 25 students in a given school and grade increases class size by 1 while moving from 25 to 26 students reduces class size by almost half (from 25 to 13). Thus, one can use the predicted class size, based on the institutional rules, as the instrumental variable for the actual class size, controlling for the number of students in the school and grade. Of courses it would be necessary to have data on actual class size in order to do this, but these data need not be attached to the NAEP dataset as one can estimate the two stages using separate datasets.

have a different set of rules governing class size. However, it might be possible to collect data on class size rules for a group of particularly large school districts and/or those covered in NAEP data, and to thereby obtain information sufficient to use class size rules in the same way they were used by Angrist, his co-authors, and Hoxby.

Internet Access

Goolsbee and Guryan (2002) estimate impacts of the E-Rate program on test scores and dropout rates in California.[23] E-Rate is a program that provides over $2 billion per year to help schools nation-wide provide Internet access to their students. The instrumental variable they use in their analysis is based on the fact that the E-Rate funding is given out based on a formula that is a very non-linear function of school poverty.[24] A similar method could be used to estimate impacts of E-Rate on NAEP scores nation-wide.

Comparison Issues

A number of issues come to mind when comparing the methods discussed above. These include whether results are biased by omitted variables, whether they are biased by measurement error in the education policy variables being considered, whether the resulting estimates are precise enough to be useful for policy purposes, and whether the relevant data are available. Following are some preliminary thoughts in each of these areas.

[23] I used a similar method in research on the E-Rate program (Puma, Chaplin, Olson and Pandjiris, 2002).

[24] Poverty is measured by the % of students eligible for free or reduced price lunch.

Bias Caused by Omitted Variables

The difference methods discussed earlier can each control for a substantial amount of bias—and in particular for any bias caused by unobserved factors that remain fixed across the sub-groups being considered. This can make these methods quite powerful. In many cases, they may be preferable to Instrumental Variables, as the assumption that the instrumental variable is uncorrelated with the resulting error term is often far from trivial. In fact, it is often found that results based on the IV method can be quite sensitive to the choice of instrumental variables. To summarize, the choice between a difference model and an IV model is generally not clear. However, in some cases one can estimate a joint model to test to see if either method suggests that the other is incorrect.

Another important issue to consider when choosing methods is whether the chosen method allows for internal consistency checks. This is possible with IV methods if one has more than one instrumental variable per education policy variable, which is often the case.[25] One can also do an internal consistency check using the cross-sectional method I described above for estimating impacts of teacher quality. This test is described in Appendix B. Evidence that the resulting model is not internally consistent would suggest that biased estimates are more likely.

The IV method does have one fairly important weakness related to bias. This is that it is possible to estimate statistically significant impacts and find no evidence of internal inconsistencies using the IV method, but to still be at great risk of bias. This can happen if the instrumental variables have very small substantive impacts on the education policy variable in the first stage of estimation. Angrist and Kruger (1995) propose a method of dealing with this

[25] In this case one can test to see if the different instruments suggest similar results.

issue that we would consider in our analysis if the instrumental variables appear to have small substantive effects.

Bias Caused by Measurement Error in the Education Policy Variables

Some researchers use measures of education policy variables that may be very imprecise and could, consequently, result in estimated impacts that would be biased downwards.[26] This could happen if, for example, one were to measure peer effects using a school average of some variable that was measured for only a subset of students at the school, say those in the NAEP data. The cross-sectional and difference models discussed above would do nothing to deal with this form of bias. In contrast, Instrumental Variables have the extra benefit of controlling for bias caused by measurement error (in addition to adjusting for bias caused by omitted variables). Thus, when measurement error is a concern, as is often the case in education policy research, Instrumental Variables may be particularly useful.

Precision of Estimates

While getting rid of bias is important, many methods that yield unbiased results may also yield very imprecise results. Both the difference and Instrumental Variables methods discussed above are likely to reduce bias, compared to a simple cross-sectional model, but often also decrease precision (i.e. increase standard errors). Consequently, the resulting estimates may tell us little about policy impacts. Simple tests exist to see if the resulting estimates are significantly different from the cross-sectional estimates.[27] If not, then many researchers opt to use the more

[26] In the extreme, if the estimate were very imprecise, it would be effectively a random variable and, consequently, have no estimated impact on the outcome in question. For this reason, measurement error generally biases estimates downwards.

[27] This can be done using a Hausman-Wu test as described in Amemiya (1985).

precise estimates and simply note that the more robust method did not suggest that the cross-sectional estimates were incorrect.

A related concern in the IV estimation is whether or not the IV predicts the education policy variable well. To check this we only need to estimate the first stage of our two-stage models. If there is no evidence of statistically significant impacts of the instrumental variable on the education policy variable, then there is no need to estimate the second stage equation, as that will also be imprecisely estimated.

One can also test the viability of some of the other ideas using a partial analysis. For instance, to estimate the impacts of class size on student outcomes one can use an instrumental variable based on the class size rules. To test this idea out it may be possible to collect data on class size regulations for a random subset of school districts, estimate the first stage of our model based on those, and then decide, based on those results, how much additional data would be needed to obtain reasonably precise estimates in the second stage.

Variation in Treatment Effects

The estimated impacts of education policies are likely to vary a great deal across individuals, over time, and for other reasons. Some of this variation may be with observed characteristics of the individuals and institutions and, consequently, be easy to incorporate into statistical models by interacting the policy variables with the relevant characteristics. Other variation is likely to be with factors that can not be observed. For this reason, the resulting estimates are often viewed as estimates of the average impact among those who were affected by the policy, and not the average impact for all individuals who could have been impacted. This problem has been investigated at some length in the econometric literature, especially with regards to Instrumental Variables. Indeed, the parameters estimated by certain types of IV

estimators are sometimes referred to as "local average treatment effects" for reasons given by Angrist and Imbens (1994). Estimates based on the other methods are also likely to represent estimated effects only for a subset of the population. A comparison across methods with regards to how generalizable the results are would be one part of this study.

Data Availability

While many instrumental variables exist in theory, they are often difficult to locate in practice. For instance, the class size instrument (the maximum class size allowed in a school) may be a variable that varies greatly across districts. Therefore, obtaining precise information may be prohibitively expensive. Similarly, in order to estimate any of the difference models described above, one would need policy variables that varied across the units being considered (i.e. over time, across cohorts, within schools, or within students). Consequently, one part of this project would be to investigate the likely availability of data needed to estimate the various models being considered.

The IV method has a distinct disadvantage with respect to data availability. In particular, it necessitates merging the instrumental variable with the NAEP data. However, it also has a potential advantage in that it does not require one to merge the education policy variable with NAEP. This is because the two stages of an IV model need not be run using the same dataset.[28] Thus, in some cases IV may enhance data availability.

---

[28] For example, in the case of E-Rate, one could estimate the association between the instrumental variable (a function of the fraction of children on free and reduced price lunch) and E-Rate money using publicly available data from the Schools and Libraries Division of the Universal Service Access Corporation (which manages the E-Rate program on behalf of the Federal Communications Commission) as that data has both the free lunch and E-Rate money

Data

In this section I describe the NAEP data, statistical issues that arise with the NAEP data, and how I address these issues.

NAEP Data

As discussed earlier, NAEP data are collected in order to assess the performance of representative samples of students at either the state or national level. Data are collected from both private and public schools. While they are not designed specifically to estimate impacts of education policy variables, many such variables do exist on the NAEP data and others can be easily merged onto the dataset. For example, the main NAEP data have included information on teachers since 1984 for both the 4th and 8th grade students and for multiple subjects. The State Assessment NAEP, started on a trial basis in 1990, also includes information on teachers. Similarly, using school IDs one can merge in information on state and district policies from other datasets.

Student Achievement: The NAEP tests cover a large number of subject areas including math, science, reading, writing, civics, world geography, U.S. history, social studies, and the arts. Each student spends about an hour filling out both the background information and doing the subject test, and each student is given a different portion of the subject test questions being given in that administration of the test.

---

information. Then, in a separate analysis one could estimate the relationship between the instrumental variable and NAEP scores. Results from these two steps can be combined to create the IV estimate without ever merging the E-Rate money data onto NAEP, which would be difficult as the E-Rate records do not all contain Ids that would enable easy merging with NAEP.

The National Assessment Governing Board (NAGB) has been in charge of selecting the subjects assessed by NAEP since 1988. Test content is based on input from teachers, curriculum specialists, subject-matter specialists, school administrators, parents, and members of the general public. According to Calderone et al. (1999), the NAEP tests "…capture a range of subject-specific content and thinking skills that students need to deal with and the complex issues they encounter inside and outside their classrooms. Furthermore, the consensual process used to develop the frameworks ensures that they are appropriate for current educational requirements." While NAEP test scores may not be the ideal outcomes for measuring impacts of many educational policies, they are explicitly mentioned in the new No Child Left Behind act and some states, such as North Carolina, have explicitly used NAEP to help design their own state assessments. Consequently, knowing the impacts of education policies on NAEP scores is likely to be of great interest to both researchers and policy makers.

NAEP Statistical Issues

There are a number of important issues one needs to consider when analyzing NAEP data. First, individual students take only a part of the NAEP exam in any subject (Lee et al., 1997). Second, each teacher/school/district/state is matched to multiple students. Third, students have differential probabilities of being in the sample. Fourth, the material assessed varies across years in the NAEP data. I deal with these issues using methods described below.

Partial Testing: In order to reduce the costs of collecting data, each student only takes a subset of the questions for each subject.[29] Individual scores for the total exam are then imputed

[29] More precisely, NAEP uses a spiraled balanced incomplete block (BIB) design. This reduces the amount of time per student and, therefore, increases the likelihood that schools will agree to participate in NAEP.

using both the student's own performance (adjusted for the part of the test that they took) and the performance of similar students. Item Response Theory (IRT) is used to adjust the student's own score for the part of the exam they took. The Average Response Method (ARM) is used to help predict the student's individual score based on their individual characteristics.[30] Finally, each student is assigned a number of "plausible values" for their test score (5 in the 1990 mathematics assessment). Thus, I have one observation for each plausible value. In order to use these multiple scores for each student, I use controls for clustering within multiple observations for a given student in STATA. This allows me to estimate the Instrumental Variable models.[31]

Multiple Students per Teacher/School/State: For a given subject, I have multiple observations (plausible values) for each student and multiple students per teacher, school, and state, depending on the analysis. Consequently it is important, in general, to control for at least two types of clustering—the within student clustering and the clustering within the units that the education policy variables vary across. Fortunately, STATA allows one to controls for two types of clustering simultaneously.[32]

Differential Sampling Probability: NAEP is designed to provide accurate estimates of the performance of students nationwide (or at the state level in the State Assessment NAEP).

---

[30] Individual scores were calculated in part based on each student's "demographic characteristics, personal attitudes and behaviors, academic behaviors and treatments, and school characteristics." (Lee et al, 1997).

[31] Some alternatives, such as HLM2PV, that are designed more specifically for the NAEP data (Lee et al., 1997), do not include IV estimation.

[32] In previous research I have found that controlling for additional levels of clustering does not generally affect the standard errors of the resulting estimates (Chaplin & Hannaway, 1998).

Students are randomly selected within schools, and schools are randomly selected, in proportion to the size of the school. However, in order to facilitate analyses of certain subgroups of the population NAEP over-samples certain types of schools. For instance, in the main NAEP, nonpublic schools and those with large minority populations are over sampled. In addition, not all selected schools are willing to participate, creating some level of non-response. For both of these reasons, weights are normally used to obtain estimates that are nationally representative of the student body (or representative for a given state for the State NAEP assessments).[33] I estimate proposed models with and without weights, adjust standard errors in the models with weights,[34] and test for differences in estimated effects between the weighted and unweighted models.

[33] Because I only have student weights, the weighted estimates are representative for students, on average, rather than for teachers or schools.

[34] The weighting causes the data to be heteroskedastic. I use robust standard errors in STATA to adjust for this and test to see if the weights matter using a test proposed by DuMochel and Duncan (1983).

References

Agodini, Robert and Mark Dynarski (2001) "Are Experiments the Only Option? A Look at Dropout Prevention Programs," Mathematica Policy Research Inc. Ref. No. 8723-300.

Amrein, Audrey L. & David C. Berliner (2002) "High-Stakes Testing, Uncertainty, and Student Learning," **Education Policy Analysis Archives**, 10(18).

Amemiya, Takeshi (1985), **Advanced Econometrics**, Harvard University Press, Cambridge, Ma.

Angrist, Joshua D. & Kevin Lang (2002) "How Important are Classroom Peer Effects? Evidence from Boston's METCO Program," National Bureau of Economic Research Working Paper 9263, October.

Angrist, Joshua D. & Alan B. Krueger (2001) "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," **Journal of Economic Perspectives**, 15 (4):69-85.

Angrist, Joshua D. & Victor Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement," **Quarterly Journal of Economics**, 114(2):533-575.

Angrist, Joshua D. & Alan B. Krueger (1995) "Split-Sample Instrumental Variables Estimates of the Return to Schooling," **Journal of Business and Economic Statistics**, 13(2):225-35.

Angrist, J.D. and G.W. Imbens (1994) "Identification and Estimation of Local Average Treatment Effects," **Econometrica,** March.

Bloom, Howard S., Charles Michalopoulos, Carolyn J. Hill, and Ying Lei (2002) "Can Nonexperimental Comparsion Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" Manpower Demonstration Research Corporation Working Paper on Research Methodology, June.

Calderone, J., L.M. King, and N. Horkay (1999) "The NAEP Guide: A Description of the Content and Methods of the 1997 and 1998 Assessments", National Center for Education Statistics Technical Report, NCES # 97990, Revised Edition, September.

Card, David and Alan B. Krueger (1996) "Labor Market Effects of School Quality: Theory and Evidence," in **Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success,** edited by Gary Burtless, Washington, D.C., Brookings Institute.

Chaplin, Duncan, Mark Turner, and Andreas Pape (2003) "Minimum Wages and School Enrollment of Teenagers: A Look at the 1990's," **Economics of Education Review**, 22(1):11-21, February.

Chaplin, Duncan and Jane Hannaway (1998) *African American High Scorers Project*. Submitted to the Mellon Foundation. (in collaboration with Stephanie Bell-Rose and Stephanie

Creaturo of the Mellon Foundation). "Technical Report One: Individual Background Characteristics and SAT Performance," "Technical Report Two: School and Neighborhood Factors and SAT Performance," "Technical Report Three: Student Activities, Course-Taking, School Performance, and SAT Performance," "Technical Report Four: College Competitiveness, Educational Aspirations, and SAT Performance."

Cook, Michael D. and William M. Evans (2000) "Families or Schools? Explaining the Convergence in White and Black Academic Performance," **Journal of Labor Economics**, 18(4):729-754.

DuMochel, William H. and Greg J. Duncan (1983) "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," **Journal of the American Statistical Association**, 78(383):535-543.

Fuchs, Victor R. and Diane M. Reklis (1994) "Mathematical Achievement in Eighth Grade: Interstate and Racial Differences," Working Paper Series No. 4784.

Goldberger, Arthur S. (1991) **A Course in Econometrics**, Harvard University Press, Cambridge, MA.

Goldhaber, Dan D. & Dominic J. Brewer (1999) "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement**", Educational Evaluation and Policy Analysis,** Forthcoming.

Goolsbee, Austan & Jonathan Guryan (2002) "The Impact of Internet Subsidies in Public Schools," National Bureau of Economic Research Working Paper, 9090.

Guthrie, John T., William D. Schafer, and Chun-Wei Huang (2001) "Benefits of Opportunity to Read and Balanced Reading Instruction on the NAEP," **Journal of Educational Research**, 94(3):145-162.

Haney, Walter (2001), "The Myth of the Texas Miracle in Education, **Education Analysis Policy Archives**, (8)41.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Stevn G. Rivkin (2001) "Does Peer Ability Affect Student Achievement?" National Bureau of Economic Research Working Paper 8502, October.

Heckman, J. and J. Hotz (1989) "Choosing among Alternative Non-experimental Methods for Estimating Impacts of Social Programs: The Case of Manpower Training," **Journal of the American Statistical Association**, 84:862-874.

Hoxby, Caroline M. (2000a) "The Effects of Class Size on Student Achievement: New Evidence From Population Variation," **The Quarterly Journal of Economics**, 115(4).

Hoxby, Caroline M. (2000b) "Does Competition Among Public Schools Benefit Students and Taxpayers?" **American Economic Review,** 90(5):1209-1238.

Hoxby, Caroline M. (2000c) "Peer Effects in the Classroom: Learning from Gender and Race Variation," National Bureau of Economic Research Working Paper # 7867.

Hoxby, Caroline M. (1996) "How Teacher's Unions Affect Education Production," **The Quarterly Journal of Economics**, 111(3):671-718.

Huang, G. & B. Yu (2002) "District Fiscal Policy and Student Achievement: Evidence from Combined NAEP-CCD Data," **Education Policy Analysis Archives**, 10 (38).

Lee, Valerie, Robert G. Croninger, and 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**, Summer, 19(2):99-121.

Manski, Charles F. (1993) "Identification of Endogenous Social Effects: The Reflection Problem," **Review of Economic Studies**, 60 (July):531-542.

Manski, Charles F. and Irvin Garfinkel (1991) **Evaluating Welfare and Training Programs**, edited by Charles F. Manski and Irwin Garfinkel, Cambridge, Harvard University Press.

Meyer, Robert H. (1996) "Value-Added Indicators of School Performance," Chapter 10 in **Improving America's Schools: The Role of Incentives**, ed. by Eric A. Hanushek and Dale W. Jorgenson, National Academy Press, Washington, D.C.

Neal, Derek (1997) "The Effects of Catholic Secondary Schooling on Educational Achievement," **Journal of Labor Economics**, 15(1,Part 1):S98-123.

Puma, Michael J., Duncan D. Chaplin, Kristin M. Olson, Amy C. Pandjiris (2002) "The Integrated Studies of Educational Technology: A Formative Evaluation of the E-Rate Program," The Urban Institute, October.

Olson, Lynn and Debra Viadero (2002) "Law Mandates Scientific Base for Research, **Education Week**, January 30th, 21(2):1.

Quality Counts 2000 (2000) "Who Should Teach?", **Education Week** (supplement) Jan., Vol. 19.

Quality Counts 1999 (1999) "State of the States", **Education Week** (supplement) Jan. Vol. 18.

Quality Counts 1998 (1998) "The Urban Challenge", **Education Week** (supplement) Jan. Vol. 17.

Quality Counts 1997 (1997) "A Report Card on the Condition of Education in the 50 States", **Education Week** (supplement) Jan., Vol. 16.

Raju, Nambury S., James W. Pellegrino, Meryl W. Berthenthal, Karen J. Mitchell, and Lee R. Jones, Editors (2000) **Grading the Nation's Report Card: Research from the Evaluation of NAEP**, Committee on the Evaluation of National and State Assessments of Educational Progress, National Research Council.

Sloop, S.L. (2002) "Impact of State Education Policy on Student Achievement: Evidence from the NAEP 1996 Mathematics State Assessment for Georgia and North Carolina," State Data and Research Center, Georgia Institute of Technology, http://docs.gadata.org/docushare/dscgi/ds.py/GetRepr/File-657/html

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**, 24(1): 1-27.

Viadero, Debra (2002) "Ed. Dept. Quietly Funds More Experimental Studies," **Education Week**, December 11th, 22(15):1.

Wenglinsky, Harold (2002) "How Schools Matter: The Link Between Teacher Classroom Practices and Student Academic Performance," **Education Policy Analysis Archives**, 10(12).

Wenglinsky, Harold (1998) "Finance Equalization and Within-School Equity: The Relationship between Education Spending and the Social Distribution of Achievement," **Educational Evaluation and Policy Analysis**, 20(4):269-283.

Wilde, Elizabeth Ty (2002) "How Close is Close Enough? Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes," Institute for Research on Poverty, Discussion Paper no. 1242-02.

Yang, Jung-Ho (2002) "Comparisons of NAEP mathematics Score Trends and Patterns for SSI and non-SSI States," paper presented at the American Educational Association annual meeting.

Appendix B: Teacher Qualifications Estimation Method

In this appendix I describe a new method that could be used to estimate the impacts of teacher qualifications on student test scores using NAEP data. I begin with a brief discussion of the teacher qualifications variables available in NAEP that could be used with this method.

Teacher Qualifications: The NAEP data include a large number of variables describing pre-service and in-service teacher qualifications. Some of these vary across the different subjects and years of the NAEP data. However, I will have a number of important variables available for most of the years and subjects including information on certification (level and subject area), academic degrees (Masters or Bachelors) and subject area of degree. In-service information includes experience (total years taught and years teaching subject) and in-service training. Additional information, available for subsets of the years and subjects, includes teacher characteristics (gender and race/ethnicity), professional development outside of subject area (for instance in technology use, cross-cultural issues, and classroom management), knowledge of specific concepts, and whether the teacher was prepared to use national standards.

Analysis Method: The proposed method is based on the idea that in many large schools more than one teacher is needed for a given subject, and, when not tracked, students are presumably allocated between these teachers based on largely non-academic factors. To the extent that this is true, I can use within-school variation to estimate the effects of teacher qualifications on student achievement.

The NAEP data contain two questions on tracking; one to the teachers, asking if their class is tracked, and a second to school administrators, asking if classes in this subject are tracked. Using this information, I will select students in schools where both the administrator

and teacher report that students are not tracked in this subject, and where I have data on students from more than one teacher of the same subject (and class). I will then run one regression for each subject (math, science, reading, and writing) where the outcomes are the student test scores, and the key independent variables are the teacher's qualifications.

The key set of controls in this analysis will be a complete set of school dummy variables—one for each school (with the intercept omitted). By including these dummies in the set of control variables I will be focusing analysis only on the differences in student test scores between teachers in the same school who are teaching the same subject and class.[35]

Specification Test for Internal Consistency: The regression method described above may not be adequate to control for differential placement of students between different types of teachers. To test for bias caused by differential placement across these "untracked" classes, I will estimate the effects of student characteristics[36] (race, ethnicity, gender, parent education, time in U.S., language spoken at home, reading material in home, and times changed schools in the past 2 years) on teacher qualifications (degree in subject, years of experience, if has a Masters Degree or beyond). In these models I will look for evidence that the student's background characteristics affect which teachers they are placed with, even though the classes

[35] We suspect that very few schools will have classes in more than one subject where each subject has multiple teachers. For 8th grade, however, we will have information on what class a student is taking in math and in science. Therefore, if we do identify such schools we will include additional dummy variables (one for each subject within the school).

[36] Some of these student characteristics are likely to be reported inaccurately, especially by the 4th graders. This will not matter if the method works.

are supposedly not tracked. Evidence of an "effect" of these student characteristics on teacher qualifications would suggest that I was not able to fully control for prior academic achievement using the sample reductions described above.[37] Evidence of this sort of bias would not make the method unusable as I can include these background factors as controls in the analyses of student test scores. Such evidence would, however, suggest that greater caution should be used when interpreting the results.

Student test scores depend on their entire schooling history, as well as on what happens to them outside of the classroom, and not just on the characteristics of the teacher they are currently with. However, if students are not being tracked, then students from different classrooms for a single subject in a single school should have similar backgrounds, on average. Any remaining differences in test performance that are systematically associated with their current teacher's characteristics are likely related to causal effects of these teacher characteristics.

In previous work I have attempted to use this method to estimate effects of teacher qualifications on 10th grade math test scores using data from the National Educational Longitudinal Survey (NELS). I identified about 2,200 students who fit the criterion described above out of a total of about 5,200 students who were taking math classes in NELS.[38] The results were not statistically significant, but the standard errors were only moderately large and

[37] We will conduct a joint significance test to see if these models, as a group, suggest evidence of tracking. It would be tempting to do a separate tracking estimate for each school but the sample sizes would probably be too small to produce reliable estimates.

[38] This is based on students with 1st follow-up test scores and non-missing values for the relevant variables.

the t-statistics were around 1. In the NAEP data I expect to have over 28,000 8th grade students in math alone, or about 12,000 students fitting the requirements for this method. This should reduce the standard errors by a factor of about 2.3, compared to using NELS.[39] Combining across grades (and testing for interactions with grade level) will increase the sample size further.[40] In addition, I expect larger sample sizes for 4th grade reading and writing.

As mentioned earlier, previous research has relied largely on controls for previous test scores to control for bias in estimated effects. If teacher qualifications are correlated with academic potential, even after controlling for previous test scores, then these standard methods may be producing biased estimates. For this reason, the method proposed here may produce more reliable estimates of the effects of teacher quality on student achievement than have been produced using the standard "value-added" methods that control for lagged test scores.[41]

Time with Teacher: The students in the NAEP sample have not necessarily been with the teacher or school they are identified to have for a full school year because the NAEP exam is

---

[39] 2.3 is the square root of 5.5, which is approximately the ratio of 29,000 divided by 5,200.

[40] We will also explore the possibility of combining across subjects, especially for reading and writing. The power of this combination is greater because separate groups of students take each subject. Finally, Al Rogers at ETS mentioned that we may be able to obtain additional data from a sample designed to focus more on LEP and Special Education students starting in 1994. This could increase those sample sizes by 30 to 50%.

[41] Ideally, we would use one data set to test both methods against each other. NELS lacks the sample size to conduct a reasonably powerful test and we are aware of no other data which could be used for this purpose.

given between January and March.[42] Fortunately, the data include the month in which the student takes the exam. Therefore, I can adjust for this by multiplying the education policy variables by the number of months the student was with the teacher.[43] Thus, the coefficient estimates will represent the estimated effects of being with that teacher for one month.[44]

---

[42] The state assessments are given in February.

[43] Assuming that they started with the teacher at the beginning of the academic year.

[44] The estimated effects may be biased downwards if some students are taking courses that did not start at the beginning of the academic year. We suspect that this is uncommon in 4th and 8th grade.