Section III.
Relating Family and Schooling Characteristics to Academic Achievement

Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88

Dominic J. Brewer and Dan D. Goldhaber
The Urban Institute

Introduction

The mission of the National Center for Education Statistics (NCES) is to collect "…statistics and information … in order to promote and accelerate the improvement of American education." To help achieve this, over the past 20 years NCES, as well as its predecessors, has greatly expanded its collection of longitudinal data. As a result, researchers have gained a better understanding of educational practices and the underlying complex relationships between students, schools, and teachers. In the absence of large-scale, randomized experiments to determine the effectiveness of educational interventions and resources, analyses of nonexperimental data can provide insights that help policymakers allocate scarce resources and enable practitioners to improve student achievement.

Analyses of NCES data have generated a large amount of literature focusing on the key determinants of student achievement and the effects of programs and policies. Educational productivity studies have focused on the overall effects of spending on schools and on the effectiveness of particular educational inputs (Monk 1992). In particular, these studies examine how per pupil expenditures and school, teacher, and class characteristics (e.g., school demographics, teacher degree levels, and class size) affect student outcomes such as test scores. This research has spawned what is commonly referred to as the "does money matter?" debate. Much of this controversy has been shaped by older studies cited in Hanushek (1986) that rely on cross-sectional and aggregated data. Although the issue is not yet settled, the availability of longitudinal data, as well as the use of more sophisticated statistical methods, has advanced our knowledge in the area. For example, earlier studies examining the impact

of class size on student outcomes used data aggregated to the school and school district levels. However, recent longitudinal data are specific enough to enable researchers to use actual class sizes rather than school or district average pupil/teacher ratios. Detailed longitudinal data have also allowed scholars to examine controversial educational reforms such as tracking and school choice.

In this paper commissioned by NCES,[1] we were asked to illustrate how we have used recent NCES longitudinal databases in our own research to investigate issues of "educational productivity," broadly defined as the relationship between school resources and educational practices and student achievement.[2] We argue that the *National Educational Longitudinal Study of 1988* (NELS:88) represents a substantive improvement over previous longitudinal data collection efforts for research and policy purposes. These enhancements permit the estimation of a wider variety of statistical models of the determinants of student achievement and allow researchers to test important hypotheses about educational practices that have implications for policy. In particular, the ability to link individual students to detailed background information about their teachers proves critical to interpreting the results of standard education production functions. As an example, our own research demonstrates (Brewer and Goldhaber 1996; Goldhaber and Brewer 1997a, 1997b) that subject-specific teacher background in mathematics and science is systematically related to student achievement in these subjects, even though teachers' higher degrees in general are not. We suggest a number of further improvements to future NCES longitudinal data studies including collection of more refined information on teacher characteristics and ability, increased frequency of follow-ups, and more student-level observations per teacher.

The paper begins with a brief introduction to the education productivity literature, followed by a discussion of some of the advantages of NELS:88 over previous NCES data, focusing on the substantive findings on teacher subject-specific preparation. Next, we argue that these data have permitted the

---

[1] This paper is based primarily on the author's research on teacher qualifications previously published in the *Journal of Human Resources, Advances in Educational Productivity, Developments in School Finance 1996,* and *Education Economics.*

[2] Hence we make no attempt to be comprehensive in reviewing other research here. That is not the purpose of this paper.

estimation of a broader set of statistical models, including hierarchical linear models, fixed and random effect models, and models with selectivity. The conclusion makes a number of suggestions for future data collection.

School Resources and Student Achievement

The ultimate reason to collect data is to influence public policy in a positive way. Thus, researchers are interested, among other things, in furnishing policymakers with the information required to make prudent resource allocation decisions and to understand which educational interventions work. Dating back to the 1966 "Coleman report" (Coleman et al.1966), there have been numerous studies on how investments in educational resources affect student performance and labor market outcomes. This line of research falls under the broad heading of "educational production functions."[3] Most educational production function studies seek to explain variance in standardized test scores at student, school, or school district levels by estimating multiple regression models that regress student outcomes on individual and family background variables and school inputs. The broad conclusion of this body of work is that individual and family traits explain the vast majority of variance in student test scores and that schools play a lesser role. Eric Hanushek notes that these studies as a whole show that "differences in [school] quality do not seem to reflect variations in expenditures, class sizes, or other commonly measured attributes of schools and teachers" (Hanushek 1986, 1142).

He concludes that there is "no strong evidence that teacher-student ratios, teacher education, or teacher experience have an expected positive effect on student achievement" and that "there appears to be no strong or systematic relationship between school expenditures and student performance" (Hanushek

[3] The notion that there is an estimable education production function for a set of individuals within or across classes or teachers or schools or school districts is not unchallenged (Monk 1992). Like any model, the education production function is certainly a simplification of reality, but it is a useful tool. This is particularly true for policy purposes because most applications focus on manipulable, measurable inputs rather than on intangible variables or amorphous constructs like "school climate" that are difficult to translate into practical recommendations.

1986, 1162).[4] Hanushek's interpretation of the literature suggests that (public) schools have a suboptimal allocation of resources (allocative inefficiency), that they do not operate on the production possibility frontier (technical inefficiency), or both. In these cases additional teacher inputs or smaller class sizes would not necessarily imply higher output, *ceteris paribus*. This result does *not* imply that schooling resources *never* affect student achievement positively, simply that, given the way public schools are organized, additional resources do not make much systematic difference.

The view that observable school inputs, and teachers in particular, do not positively impact student achievement rests on somewhat shaky empirical ground. Hanushek's conclusion is based primarily on older work, and there are good reasons to believe that much educational productivity research completed in the 1970s had major deficiencies. One problem is likely to be that key variables may have been omitted from estimated test score models, potentially leading to biased coefficient estimates of the included variables.[5] Missing information and crude proxies for many schooling inputs in older data make this likely. For example, many early studies were unable to control for prior achievement using a "pre-test" score to net out individual ability (Boardman and Murnane 1979; Hanushek 1979; Hedges, Laine, and Greenwald 1994).

Additionally, schooling inputs—notably class size and expenditures per pupil—are measured with some degree of error. This error arises, in part, from aggregation of variables to the school or district level (Hanushek, Rivkin, and Taylor 1996). For instance, rather than class size, studies often utilize total school enrollment divided by the total number of teachers (or professionals) as

---

[4] Hanushek concluded that there is no *systematic* relationship between observable schooling resources and student test scores, at least at current levels of resource utilization, by noting the direction of estimated input effects (teacher/pupil ratio, teacher education, teacher experience, teacher salary, per pupil expenditures, administrative inputs, and facilities) on student achievement, along with whether they were statistically significant, and simply tallying ("vote counting") the number of statistically significant positive and negative coefficients. A "meta-analysis" by Hedges, Laine, and Greenwald (1994) using the same set of studies reviewed by Hanushek reached a very different conclusion. Their basic argument was that the *pattern* of estimated coefficients in these studies suggested there were indeed systematic positive effects, although it is not clear that the alternative interpretation gives any clearer guidance for policymakers.

[5] Omitted variables and aggregation bias problems in the context of educational production functions are discussed in a more formal fashion in Goldhaber and Brewer (1997b); see also Hanushek (1979) and Hanushek, Rivkin, and Taylor (1996).

an average pupil/teacher ratio. This is problematic given that there is considerable variation in class size *within* schools as well as *between* schools. Variables representing school and teacher "quality" used in most production function studies are typically very crude. For example, teacher degree level and years of experience may be only weakly related to teaching skill. Degree alone does not distinguish among diplomas given at high- and low-quality colleges or when the degree was granted, nor does it convey any information about college major, certification requirements fulfilled, or subsequent professional development. Teacher motivation, enthusiasm, and skill at presenting class material influence students' achievement, but are difficult traits to accurately measure and are, thus, omitted from standard regression analyses.

Recent Longitudinal Data: Reducing Omitted Variables and Aggregation Bias

Advances in statistical techniques and the collection of two large-scale longitudinal databases by NCES, *High School and Beyond* (HS&B) and the *National Educational Longitudinal Study of 1988* (NELS:88), have allowed researchers to learn more about how school resources impact students. In this section we first review the advances in these data and introduce our work on teacher subject-matter preparation, which reduces the omitted variables and aggregation bias problems inherent in earlier studies.

HS&B and NELS:88

High School and Beyond (HS&B) was one of the first large-scale longitudinal databases. A cohort of students were tested in both tenth and twelfth grades, permitting researchers to use a "value-added" methodology—examining how much students learned between two points in time (as measured by standardized tests), using a pre- and post-test. Unfortunately, teacher data were only collected in 1984, two years after the students had graduated and, therefore, could not be linked back to particular students. Hence, like previous work, statistical models estimated using individual level test scores had to rely on school-level measures of variables such as class size and teacher experience (Ehrenberg and Brewer 1994).

The NELS:88 data represent a substantial improvement over HS&B because they include concurrent, detailed school and teacher data collected in such a way that researchers can link students to their particular teachers and

classes. The NELS:88 study is a nationally representative survey of about 24,000 eighth grade students conducted in the spring of 1988, with follow-ups conducted in 1990, 1992, and 1994. At the time of each survey, students took one or more subject-based tests in math, science, English, or history. The tests were carefully designed to avoid "floor" and "ceiling" testing effects and were put on a common scale using Item Response Theory.[6] Linked student-teacher-class data allow an investigation of the impacts of specific class size, teacher characteristics, and peer effects on student achievement. NELS:88 is constructed in such a way that students can be linked to data gathered in a separate teacher survey that provides information on the teacher's background and teaching methods and curricula used in the particular class the tested student is taking. This represents a major advance.

Of course, there are still some important deficiencies in the NELS:88 design. For example, the study skips a year in sampling students so information on the characteristics of schools, teachers, and classes is missing for the ninth and the eleventh grades. Because information on the ninth and eleventh grades is left out of the survey, studies that use the NELS:88 may suffer from omitted variables bias. The direction and magnitude of this bias depends on the relationship between the tenth and/or twelfth grade characteristics included in the survey and those in the excluded grades. Further, in each follow-up to the eighth grade base year, there are fewer students per class (i.e., there is a fanning out of the original set of eighth grade students to multiple teachers and classes by the tenth grade). As a result, in some cases there are as few as one student per class and teacher in the follow-up surveys. This makes it more difficult to distinguish between teacher and class effects.[7]

Several researchers have taken advantage of the ability to link students to their classes and teachers. For example, the magnitude of the effects of class size on student achievement has long been debated, but there is currently widespread interest in class size-reduction policies at federal and state levels (Parrish and Brewer 1998). In the absence of large-scale experimental data, such as

---

[6] For more information on this methodology, see Rock and Pollock (1991).

[7] This is not a problem for some kinds of statistical model (e.g., ordinary least squares, random effect models). In all of our work we use tenth grade school, teacher, and class variables, and in most cases impose no restrictions on the minimum number of student observations per class or per teacher.

Project STAR in Tennessee, nonexperimental production function type studies will continue to be important, so it is necessary to have data that allow for the best possible test of the relationship between class size and student outcomes.

Many of the class size studies cited by Hanushek (1986) have found no relationship between class size and student achievement. Some have even found that achievement is higher in larger classes—this is clearly counterintuitive. Using NELS:88, Akerhielm (1995) shows that students are not randomly distributed across classes within schools. Low-achieving students tend to be assigned to smaller classes. Without accounting for this nonrandom assignment of students to classes, there appears to be a positive relationship between class size and achievement. However, when statistical models incorporate this nonrandom assignment, the relationship becomes negative (i.e., smaller classes result in higher achievement) between class size and student achievement. Because NELS:88 links students to a particular class, Akerhielm was able to measure the actual class size rather than an aggregate pupil/teacher ratio.

Teacher Subject-specific Preparation

The ability to link students to their particular teachers afforded by NELS:88 has also allowed researchers to better understand how teachers affect students. Ehrenberg, Goldhaber, and Brewer (1995) were able to shed some light on the issue of "role models" in education by investigating whether the race-ethnicity and gender of teachers impact student test scores.[8] Similarly, NELS:88 permits an investigation of whether the subject-specific preparation of teachers affects student achievement. The National Commission on Teaching and America's Future in 1996 reported that one-fourth of high school teachers lack college training in their primary classroom subject. Underlying this concern about out-of-field teaching is the assumption that teachers with degrees in the subject that they teach are more effective. Although this may seem a commonsense proposition, there is relatively little quantitative work on the relationship between educational outcomes and teacher subject-specific

[8] This research shows that, on balance, teachers' race-ethnicity and gender are more likely to influence teachers' subjective evaluations of their students than to affect student achievement as measured by standardized tests. The important point here, however, is that this research would not be possible without teacher data directly tied to individual students.

preparation, because most data used for these analyses do not contain this information.[9]

Using samples of students in four subjects,[10] Goldhaber and Brewer (1997a) estimated standard education production function models in which a student's tenth grade test score in a subject is regressed on (1) individual and family background variables (e.g., sex, race-ethnicity, parental education, family structure, family income, and eighth grade test score); (2) tenth grade school level variables (e.g., urbanicity, regional dummies, school size, the percentage of students who are white, the percentage of students from single-parent families, and the percentage of teachers with at least a master's degree); (3) tenth grade teacher variables (e.g., sex, race-ethnicity, years of experience at the secondary level, whether the teacher is certified, and his or her degree level); (4) and tenth grade class-level variables (e.g., class size and percentage of minority students in the class). The results from these models demonstrate two important things. First, school level aggregates, such as the percentage of teachers in a school with at least a master's degree—the extent of information on teachers available in previous data, are statistically insignificant in all estimated statistical models regardless of subject.

Second, the ability to include subject-specific teacher degree and certification information is critical to interpreting the results of these statistical models. This is illustrated in table 1. Columns (1), (3), (5), and (7) of the table show the estimated regression coefficients of the teacher variables included in the model, while columns (2), (4), (6), and (8) show the results when we include more refined subject-specific teacher characteristics (whether the teacher is certified in his or her subject area and whether the teacher has BA or MA degrees in his or her subject area). These variables allow us to distinguish between teachers who have BA or MA majors in the subject they teach, those who have certification in the subjects they teach, and those who do not have subject-specific training. In the models reported in columns (1), (3), (5) and (7), *years of teaching experience* is not a statistically significant item, nor is whether the teacher

---

[9] Monk and King (1994) report that teacher subject matter preparation in mathematics and science does have some positive impact on student achievement in those subjects. Again, this insight is made possible only because the *Longitudinal Survey of American Youth* (LSAY) data that they used links students with individual teachers.

[10] The sample sizes for the four subjects were 5,113 students in math; 4,357 students in science; 6,196 students in English; and 2,943 students in history.

Table 1. Comparison of Selected Coefficients from Educational Production Functions[a] (absolute value of *t*-statistic)

	Math		Science		English		History	
Teacher Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Years of experience at secondary level	0.018 (1.5)	0.013 (1.1)	0.007 (0.7)	0.007 (0.6)	-0.007 (0.6)	-0.007 (0.7)	0.025 (1.6)	0.025 (1.7)
Certified	-0.511 (0.9)	-2.343 (2.3)	0.140 (0.3)	-0.827 (1.2)	-1.267 (1.9)	-0.645 (0.7)	0.170 (0.2)	0.142 (0.1)
Certified in subject	-	2.172 (2.2)	-	1.130 (1.2)	-	-0.685 (0.9)	-	0.035 (0.0)
BA or more in subject	-	0.769 (3.6)	-	0.683 (3.3)	-	0.130 (0.3)	-	-0.243 (0.8)
MA or more degree	0.247 (1.2)	0.052 (0.2)	0.030 (0.2)	0.023 (0.1)	-0.070 (0.4)	-0.085 (0.4)	-0.038 (0.1)	-0.056 (0.2)
MA or more in subject	-	0.595 (2.1)	-	0.002 (0.0)	-	0.078 (0.3)	-	0.101 (0.3)
Sample Size	5113	5113	4357	4357	6196	6196	2943	2943
Adjusted R^2	0.766	0.767	0.377	0.378	0.605	0.605	0.275	0.274

[a]Models also include the following variables: sex, race-ethnicity, parental education, family structure, family income, eighth grade test scores, urbanicity, regional dummies, school size, the percentage of students who are white, the percentage of students from single-parent families, the percentage of teachers with at least a master's degree, tenth grade teacher sex and race-ethnicity, and tenth grade class size and percentage of minority students in the class.

NOTE: With these sample sizes, any *t* statistic greater than 1.645 indicates the estimate is statistically significant at the 10 percent level, and 1.960 indicates the estimate is statistically significant at the 5 percent level.

Reproduced from Goldhaber and Brewer (1997a).

has a master's degree—implying that teachers with master's degrees are no more (or less) effective than those without advanced degrees. The results for teacher certification are similar in that we find the coefficient on teacher certification to be statistically insignificant (except in English, where teacher certification is significant and negative). If these were the only variables available, one might erroneously conclude that teacher preparation does not matter.

By contrast, in math and science, teacher subject-specific training has a statistically significant impact on student test scores in those subjects [columns (2) and (4)]. A teacher with a BA or an MA in math has a statistically significant *positive* impact on students' achievement relative to teachers with no advanced degrees or degrees in non-math subjects.[11] Further, these findings appear to reflect subject-specific training rather than simply teacher ability.[12]

Table 2 shows the estimated effects of model specification on predicted tenth grade achievement scores in math and science (we do not show English and history, because none of the subject-specific variables were statistically significant). We can infer the magnitude of the effect of teacher training on student achievement by examining the estimated coefficients in the models that include subject-specific information. For example, the effect of a teacher's having an MA in math is the sum of the coefficients of MA and MA major in math. We see the impact of model specification in math and science by comparing columns (1) and (2) for math and columns (3) and (4) for science. In both math and science, a subject-specific BA degree improves student achievement; and the results are even more pronounced for math, where an MA in math also has a statistically significant effect on achievement. In the model with general teacher variables, we predict students (with average characteristics) who have a teacher who is certified in math and has both a BA and an MA in math to have a tenth grade math score of 44.06. However, these same stu-

---

[11] See Goldhaber and Brewer (1997a, 1997b) for more details on various robustness checks and statistical tests performed on these models. We find no evidence that certification or subject-specific degrees have an effect on student achievement in English or history, where the subject-specific variables were statistically insignificant.

[12] Teacher math and science degrees may serve as proxies for teacher ability. To test this hypothesis, we re-estimated all models, including whether a teacher has a math or science degree in the English and history regressions. If math and science degrees serve as proxies for teacher quality, we would expect the coefficients on these variables to be significant and positive in all of the subject areas, including English and history. This is not the case. Neither the math nor the science degree level variables are statistically significant in the English and history regressions.

Table 2. Effect of Model Specification on Predicted Test Scores[a]

	Math		Science	
	Model with general teacher variables	Model with subject-specific variables	Model with general teacher variables	Model with subject-specific variables
Certification (in subject)	**43.94**	**43.95***	**21.79**	**21.81**
BA only (in subject)	**43.96**	**44.21***	**21.78**	**21.99***
MA only (in subject)	**44.08**	**44.57***	**21.79**	**21.78**
BA, MA, and certification (in subject)	**44.06**	**44.69***	**21.80**	**22.02**

[a]All other variables are measured at their mean value.

* Indicates statistically significant difference from base model with general teacher variables.

Reproduced from Goldhaber and Brewer (1997a).

dents are predicted to have a tenth grade math score of 44.69 when the subject-specific specification of the model is used. The difference between these predicted scores, .63, is about 5 percent of the tenth grade math test standard deviation, a relatively small (but statistically significant) difference. This finding is important, because it suggests that student achievement in technical subjects can be improved by requiring more in-subject teaching. The estimation of statistical models that yield this finding is only made possible by the linked and detailed data available in NELS:88.

Recent Longitudinal Data: Permitting a Broader Class of Statistical Models

In addition to reducing the omitted variables and aggregation biases in earlier production function studies problems (Goldhaber and Brewer 1997b; Hanushek, Rivkin, and Taylor 1996), the NELS:88 data allow researchers to estimate a broader class of statistical models that may have useful implications for understanding educational productivity.

Hierarchical Linear Models

One example of such a model type made possible by NELS:88 is illustrated by Lee and Smith (1997), who investigated the relationship between school size and student achievement using a technique called hierarchical linear modeling (HLM). The basic approach of HLM is to first estimate a within-group model and use the estimated slope coefficients as the dependent variable in a second across-group stage. The HLM technique has become widely used by educational researchers to model effects that are thought to correspond to particular levels or groupings. It may be argued that standard statistical models (e.g., ordinary least squares) yield inefficient estimates of the effects of some schooling variables when those variables affect groups of students jointly. This is because individual level models, which include higher-level effects (such as school climate), may not adequately capture the contextual impact of these higher-level effects. Thus, the NELS:88 data permit these types of models because they contain multiple observations per class, teacher, and school. Several researchers have taken advantage of this data structure to examine different hypotheses.[13]

---

[13] For instance, see Lee, Smith, and Croninger (1997) for an investigation of how the social and academic organization of high schools affect learning in math and science and Gamoran (1996) for a comparison of public and private school achievement.

Although HLM is intuitively appealing and may yield efficiency gains, it is important to note that, like any other statistical model, it requires a particular data structure and is predicated on a set of assumptions, such as the distribution of the error term. Further, it is only possible to estimate these models when there are multiple observations at each contextual level.[14]

Thus, with the NELS:88 data, HLM utilizes only a sub-sample of all the potential students in the sample and some information is lost. Additionally, the choice over the level of the effect is somewhat arbitrary. For instance, Lee and Smith (1997) specify the effect at the school level but ignore the potential for class-level effects. Similarly, one might argue that the contextual effect is at peer (small group) or neighborhood levels, both of which are typically ignored. Finally, HLM does not permit researchers to handle situations when the first-level outcomes and second-level regressors are jointly determined (endogenous) (Mason 1995). For instance, school quality may play an important role in influencing parental choice of school sector. Unless models explicitly account for this type of selection, they will yield biased coefficient estimates. The bottom line is that although HLM may yield more efficient estimates of contextual effects, it only addresses one of the many problems associated with using nonexperimental data, such as omitted variables bias, sample selection bias, and endogeneity.

Fixed- and Random-Effects Models

The NELS:88 data do allow researchers better opportunities to investigate some of these issues. For example, Goldhaber and Brewer (1997b) use the data to test whether unobservable teacher characteristics cause systematic bias in the estimated effects of observable variables. This is made possible because the structure of the data permits the estimation of fixed- and random-effects models. The results of this work suggest that unobservable teacher characteristics such as motivation and skill are not systematically related to observable teacher characteristics and that this does not cause bias in standard educational production function studies.

In follow-up work, Goldhaber, Brewer, and Anderson (1999) calculate the role of individual, family, school, teacher, and class characteristics in ex-

[14] It is not clear, however, how many observations are required at each level to adequately capture the contextual effects in question.

plaining variance in student (math) test scores. In doing so, they distinguish between the contributions of observable and unobservable factors. They find the vast majority of variance is explained by individual and family background characteristics (about 60 percent). Overall, school, teacher, and class variables, both observed (e.g., class size and teacher certification and degree level) and unobserved (e.g., teacher motivation and parental involvement), account for approximately 21 percent of the variance in student achievement. Of this 21 percent, only about 1 percentage point (or 4.8 percent) is explained by observable educational variables, and the remaining 20 percentage points (or 95.2 percent) are made up of unobservable school, teacher, and class effects.

These unobservable effects may represent variables, such as the climate in the school, the students' peers, the ability of the teacher, or unobservable family background characteristics that are not adequately controlled for in the model.

Selectivity-Corrected Models

The issue of student selection also often arises when researchers attempt to understand how schools impact students. Student selection is another form of omitted variables bias that occurs when individual characteristics not easily quantified in data are associated with a choice made by those individuals. For instance, one educational reform that has recently gained a great deal of attention is the use of educational vouchers (school choice). Proponents of choice often point to student success in private schools as evidence of greater educational productivity in the private sector.[15] However, private schools can establish admissions criteria, such as minimum test scores, whereas public schools, in general, must accept all students. Private schools also tend to serve students whose parents are more affluent and well educated. Thus, it is not immediately clear that differences in performance between public school students and private school students are a direct result of the delivery of education or due to differences in their background.

Numerous studies have examined this issue, notably an HS&B-based study by Coleman et al. in 1981. This work was widely criticized because it did not

[15] On average, private schools have higher standardized test scores, graduation rates, and college matriculation rates than do public schools.

adequately control for selection into school sector. Goldhaber (1996) and Figlio and Stone (1997) both use the NELS:88 data to tackle the selection issue in the context of public and private schools.[16] They find little or no difference in the effectiveness of public and private schools. Simulations of what would happen if a student were to switch school sectors indicate that, holding all else equal, the large difference between public and private schools in mean standardized test scores is accounted for primarily by student background, school resources, and student selection, rather than differences in how effectively schools use the resources that they have.

Other data would permit a simulation of this type but would not account for teacher- and class-level differences. A very similar procedure has also been used in a series of studies examining the effectiveness of ability grouping (tracking).[17]

Conclusion: Further Improvements in the Data

In this paper we have argued that recent improvements in longitudinal data collection permit researchers to better tackle important unresolved educational policy issues, such as the effects of class size and teacher preparation on student achievement and the effectiveness of private schools. The availability of more detailed data with students linked to teachers reduces the likelihood of aggregation and omitted variables biases, and the design of NELS:88 with student-, class-, teacher-, and school-level information permits the estimation of a broader class of statistical models than has previously been possible. Although large-scale controlled experiments are clearly preferable, we believe that the advances made in nonexperimental data collection and methodology over the past decade represent a substantive improvement.

---

[16] This methodology uses a statistical procedure known as the Heckman two-step method (Heckman 1979), which requires variables that affect choice of school sector but do not affect student achievement. NELS:88 contains a richer set of variables that potentially fulfill this requirement. See Figlio and Stone (1997) for a detailed discussion of this point.

[17] Although a number of small-scale experiments have suggested that placing students in classes of heterogeneous ability benefits all students, this conclusion is controversial. Large-scale nonexperimental studies using NELS:88 suggest the effects are not so simple. See, for example, Brewer et al. (1995) and Argys et al. (1996).

We believe a number of improvements could be made in future NCES longitudinal data to further reduce the potential for omitted variables and aggregation biases and allow for the estimation of more sophisticated statistical models. For example, one problem with NELS:88 is that students are only surveyed every two years. This means that data about the students' academic and other experiences in the intervening year are lost.[18] As noted above, this creates the potential for bias in any model of student achievement growth. Additionally, it would be useful for research purposes to have more detailed class-level information, particularly on the socioeconomic status of the students in each class. This is crucial to understanding peer effects. The ideal would be to continue to collect data on teachers and classes and test scores that correspond to the students' exposure to particular teachers and classes.

As we have stressed in this paper, the link between students and teachers is a critical addition to NELS:88 and should be maintained. However, this link would be even more useful if additional data about teacher background were available. For instance, the few studies that have had measures of teacher (verbal) ability, for example, in the form of a teacher test score or selectivity of undergraduate institution, have found it is related to student achievement (Coleman et al.1966; Ehrenberg and Brewer 1995; Ferguson 1991). However, collecting a teacher test score could be controversial. It would be necessary to obtain information on standardized tests that teachers have taken in the past (e.g., SAT, ACT, and NTE) either through transcript collection or by administering new tests that measure teacher knowledge and ability. Either option could be costly and time-consuming. A third alternative that is likely to be useful, but less costly, is to collect evaluations of individual teachers by the school principal (since the principal is already being surveyed).

Information on the year that teachers obtained their licensure and the state from which they obtained their licenses would also be quite useful. This would be a relatively low-cost addition—perhaps as little as one item on the teacher survey—but it would allow researchers to better identify the effects of state and institutional policies. This appears particularly important given that policymakers in many states have recently overhauled (or are considering changing) their licensure and/or teacher preparation requirements.[19]

---

[18] Given budgetary constraints, one option would be to collect an abbreviated set of student and teacher variables in the intervening years.

[19] One possibility is to collect information in such a way that future data can be linked to the *Schools and Staffing Survey*, which contains these types of data.

In considering possible additions to the data, we recognize that there are constraints on the amount of information that can be collected. Thus, we would de-emphasize the importance of collecting items relating to student, parent, and teacher beliefs, attitudes, and feelings. We consider these to be intermediate variables that are, in many cases, byproducts of actual individual qualities and institutional policies. The rationale for this suggestion is that policymakers can only indirectly influence beliefs, attitudes, and feelings through the incentive structures they create (e.g., the structure of teacher compensation) through policy levers.

A major cost saving could also be achieved by putting less stress on collecting nationally representative samples. Sampling fewer schools with more data on students and classes within a smaller number of schools is an alternative.

We recognize that NCES has an obligation to provide national educational indicators; however, this is less important for the kinds of multivariate statistical models that researchers find most persuasive in tackling the most important policy questions. The reason is that, for statistical purposes, it is not necessary to have a nationally representative sample to obtain accurate estimates of the effects of the variables of interest.

Finally, one weakness of NELS:88 is the limited number of student observations per teacher and class. This limitation means that it is difficult to separate teacher effects from class-level effects. The more student observations per class and the more classes per teacher, the more we can learn about how teacher characteristics or behavior affects student outcomes and how these factors are related to the types of students being taught. While our findings on teacher preparation derived from the NELS:88 study are important for policy, future data collection could allow researchers to gain a more comprehensive understanding of the complex relationships between students, teachers, and schools.

References

Akerhielm, K. (1995). Does Class Size Matter? *Economics of Education Review 14*(3): 229–241.

Argys, L., Rees, D. I., and Brewer, D. J. (1996, fall). Detracking America's Schools: Equity at Zero Cost. *Journal of Policy Analysis and Management 15*(4).

Boardman, A. E., and Murnane, R. J. (1979). Using Panel Data to Improve Estimates of the Determinants of Educational Achievement. *Sociology of Education 52*: 113–121.

Brewer, D. J., Rees, D. I., and Argys, L. M. (1995, November). Detracking America's Schools: The Reform Without Cost? *Phi Delta Kappan 77*(3): 210–215.

Brewer, D. J., and Goldhaber, D. D. (1996). Educational Achievement and Teacher Qualifications: New Evidence from Microlevel Data. In B. Cooper and S. Speakman (Eds.), *Optimizing Education Resources* (pp. 389–410). Greenwich, CT: JAI Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity.* U.S. Department of Health, Education and Welfare. Washington, DC: U.S. Government Printing Office.

Coleman, J. S., Hoffer, T., and Kilgore, S. (1981). *Public and Private Schools*. *Report to NCES.* Chicago: National Opinion Research Center.

Ehrenberg, R. G., and Brewer, D. J. (1994). Do School and Teacher Characteristics Matter? Evidence from *High School and Beyond. Economics of Education Review 13*(1): 1–17.

Ehrenberg, R. G., and Brewer, D. J. (1995). Did Teachers' Verbal Ability and Race Matter in the 1960s? *Coleman* Revisited. *Economics of Education Review 14*(1): 1–23.

Ehrenberg, R. G., Goldhaber, D. D., and Brewer, D. J. (1995). Do Teachers' Race, Gender, and Ethnicity Matter? Evidence from NELS:88. *Industrial and Labor Relations Review 48*(3): 547–561.

Ferguson, R. (1991). Paying for Public Education: New Evidence on How and Why Money Matters. *Harvard Journal on Legislation 28*: 465–498.

Figlio, D., and Stone, J. (1997). School Choice and Student Performance: Are Private Schools Really Better? Unpublished paper. Department of Economics, University of Oregon.

Gamoran, A. (1996). Student Achievement in Public Magnet, Public Comprehensive, and Private City High Schools. *Educational Evaluation and Policy Analysis 18*(1): 1–18.

Goldhaber, D. D. (1996). Public and Private High Schools: Is School Choice an Answer to the Productivity Problem? *Economics of Education Review 15*(2): 93–109.

Goldhaber, D. D., and Brewer, D. J. (1997a). Evaluating the Effect of Teacher Degree Level on Educational Performance. In W. Fowler (Ed.), *Developments in School Finance, 1996* (NCES 97–535) (pp. 199–208). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Goldhaber, D. D., and Brewer, D. J. (1997b). Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. *Journal of Human Resources 32*(3): 505–523.

Goldhaber, D. D., Brewer, D. J., and Anderson, D. (1999). A Three-Way Error Components Analysis of Educational Productivity. *Education Economics 7*(3): 199–208.

Hanushek, E. A. (1979). Conceptual and Empirical Issues in the Estimation of Education Production Functions. *Journal of Human Resources 14*(3): 351–388.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in the Public Schools. *Journal of Economic Literature XXIV*(3): 1141–1178.

Hanushek, E. A., Rivkin, S., and Taylor, L. (1996). Aggregation and the Estimated Effects of School Resources. *Review of Economics and Statistics 78*: 611–627.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1): 153–161.

Hedges, L.V., Laine, R., and Greenwald, R. (1994). A Meta-analysis of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher 23*(3): 5–14.

Lee, V. E., and Smith, J. (1997). High School Size: Which Works Best and for Whom? *Educational Evaluation and Policy Analysis 19*(3): 205–227.

Lee, V. E., Smith, J., and Croninger, R. (1997). How High School Organization Influences the Equitable Distribution of Learning in Mathematics and Science. *Sociology of Education 70*: 128–150.

Mason, W. M. (1995). Comment. *Journal of Educational and Behavioral Statistics 20*(2): 221–227.

Monk, D. H. (1992). Educational Productivity Research: An Update and Assessment of Its Role in Education Finance Reform. *Educational Evaluation and Policy Analysis 14*: 307–332.

Monk, D. H., and King, J. (1994). Multi-level Teacher Resource Effects on Pupil Performance in Secondary Mathematics and Science: The Role of Teacher Subject Matter Preparation. In R. Ehrenberg (Ed.), *Contemporary Policy Issues: Choices and Consequences in Education.* Ithaca, NY: ILR Press.

Parrish, T., and Brewer, D. J. (1998, August). *The Costs of Enrollment Increases and Class Size Reduction.* Report to U.S. Congress. Palo Alto, CA: American Institutes for Research.

Rock, D. A., and Pollock, J. M. (1991). *Psychometric Report for NELS:88 Base Year Test Battery.* Washington, DC: National Center for Education Statistics.

School-level Correlates of Reading and Mathematics Achievement in Public Schools

Donald McLaughlin and Gili Drori
American Institutes for Research

Introduction

The Schools and Staffing Survey (SASS) conducted by the National Center for Education Statistics (NCES) offers the most comprehensive picture available of the education system in the United States. Initiated in 1987–88 and repeated in 1990–91 and 1993–94, SASS consists of surveys of districts, schools, principals, and teachers that are associated with a national sample of schools. It offers information on issues such as policies, programs, services, staffing, and enrollment at both the district and the school levels, as well as the principals' and teachers' background, training, experience, perceptions, and attitudes. Given the broad reach of SASS, it can speak to a variety of important educational research and policy questions. The value of SASS would be even greater, however, if the relationship between these measures and the level of achievement in schools were known. As noted by others (Boruch and Terhanian 1996, Kaufman 1996), combining this survey information with data from other sources would allow SASS to more meaningfully inform debates over which factors relate to school effectiveness and could contribute to a broad-based evaluation of school improvement strategies.

The aim of this paper is to show the potential value of a linkage between the SASS database and information on student academic achievement collected

by individual states.[1] Most states currently collect state assessment data[2] on their public schools and thus offer state-specific information on school performance in terms of student test scores. Although many different assessment instruments are used across the states, they all aim to provide an indication of the reading and mathematics achievement levels of their schools. By transforming each school's score to a *z score* relative to other schools tested at the same grade in a state, there is potential for pooling analytical results across states to increase both power and generalizability.

While pooling information from individual states into a single database can add substantial power to analyses identifying school-level correlates of achievement in SASS schools, it does not capture between-state sources of covariation with achievement. State policies and state demographics frequently limit the variation of education practices in a state, so that within-state associations with achievement are attenuated. To capture the full range of achievement variation between schools across the nation, one must include between-state variation. That variation is reported by State NAEP, the component of the National Assessment of Educational Progress (NAEP) which focuses on state-by-state achievement assessment. To evaluate whether combining state assessment scores with State NAEP data would increase the value

---

[1] This paper is based in part on a project carried out by the authors for the National Center for Education Statistics, through the Education Statistics Services Institute. The authors wish to thank the project officer for that project, J. Michael Ross, for his thoughtful suggestions and encouragement during the course of the work. We also wish to thank Adam Gamoran, Robert Hauser, Valerie Lee, and Stephen Raudenbush for their thoughtful reviews of an earlier draft. Nevertheless, the conclusions expressed in this paper are solely those of the authors, and no endorsement of these conclusions by reviewers or by the Center should be inferred.

We appreciate the help provided by the Center and 20 State Education Agencies in providing the data for this project, and we appreciate the work of staff of the American Institutes for Research, including Mary Anne Arcilla, Grace Wu, Elizabeth Hartka, and Inna Shapotina, in putting the database together. Finally, we recognize the thousands of hours of time spent by the respondents to these surveys and assessments, who provided information that can be used to improve education policies and practices.

[2] In 1994–95, 45 U.S. states had a statewide assessment system; the remaining five states either did not have a statewide system at all or had temporarily suspended their programs (National Education Goals Panel 1996). In 1995–96, 46 states administered statewide assessments (Roeber, Bond, and Braskamp 1997). In 1996–97, 45 states administered statewide assessments (Roeber, Bond, and Connealy 1998). Some educational assessment is done in every state, and in most of the few states without statewide testing programs, most districts use nationally standardized tests for assessment.

of the linked database, this study focused on states which both conducted statewide assessments in 1993–94 and participated in the 1994 State NAEP fourth grade reading assessment. Thus, in each of the 20 states included in this study, individual state assessment data are available for most or all public schools, along with summary NAEP results based on approximately 2,500 students in 100 random schools in the state.

To assess the potential value of combining SASS with achievement data at the school level requires three steps: (1) matching the schools on the 1993–94 SASS file with state reading and mathematics assessment scores for public schools in 20 states; (2) creating a comparable achievement measure for the matched schools from the combination of state assessment and State NAEP information; and (3) carrying out analyses to test hypotheses by modeling the relationship between a variety of SASS school-level responses and average student assessment scores at the school level. The hypotheses selected for the third step concern the identification of school-level correlates of student achievement. Although analyses of school-level information collected at one point in time cannot be used either to identify individual-level correlates of achievement or to draw direct causal inferences, they can potentially provide a basis of evidence for addressing issues of strategies for school improvement.

Combining SASS and State Assessment Data

The linkage of state assessment data to the SASS file required access to restricted information concerning the identification of SASS schools. NCES has established clear criteria for acceptable procedures for storing and using confidential information, and the American Institutes for Research (AIR) have complied with these criteria. Although the linkage might have been possible with information about schools' names and addresses, it was greatly facilitated by the use of an intermediate linkage of both SASS and state assessments through the Common Core of Data (CCD). The 1991–92 CCD file, which served as the sampling frame for SASS, identifies most of the 86,287 public schools in the country by both their federal and their state identification codes.[3]

[3] Information about CCD can be found on the NCES Web page: http://www.ed.gov/NCES/.

In most cases, school records on state assessment files included the state's identification code, which enabled an automated matching procedure.[4]

These 20 states contained 3,785 of the 8,767 SASS public schools. Of these 3,785 SASS public schools, 2,916 had students enrolled in grades corresponding to the state assessment; and 2,628 were identified as having both SASS and state assessment information.[5] Of these, 66 had no teacher data, and one had erroneous mean achievement scores, so the final file used for analysis contained 2,561 school records: 1,123 elementary schools, 496 middle schools, 595 secondary schools, and 347 combined-grade schools. The database includes at least 50 schools in each state and constitutes a broad sample of large and small, urban and rural, affluent and impoverished public schools.

The coverage of the range of educational contexts in the United States by the schools in the sample determines the extent to which inferences based on analyses of the database can be expected to generalize to other schools in the country. Although SASS includes both public and private schools, state assessment data are collected for public schools only; hence, the SASS student achievement subfile created for this report is limited to information on public schools.

The 2,561 public schools included in the study are only slightly different from the general population of American public schools on most measures examined. Although 51 percent of the elementary, middle, and secondary schools in the study sample were elementary schools, compared to 63 percent in

---

[4] A few of the schools were identified "manually" by matching their state, city, or zip code, either because the federal identification code was missing from the restricted SASS file or because the state identification code was not included on the assessment file. Details of the file development process can be found in Wu, Royal, and McLaughlin (1997).

[5] Of the other 288 SASS schools, 254 did not match with state assessment, and 34 merged with state assessment files but did not have usable mean scores in both reading and mathematics. In addition, 112 of the 254 nonmatching schools were special, alternative, or vocational education schools, or schools that had an enrollment of fewer then 10 in the grade assessed; and 86 were not included in one state's (Pennsylvania's) assessment sample in 1993–94. As a measure of the success of the matching process, 2,662 of the 2,718 SASS public schools that were expected to have matching assessment scores were matched, for a match rate of 97.9 percent.

the nation[6], the percentages of central city, urban fringe, and rural schools in the study sample were each within one percent of the percentages in the population.[7]

Results of analyses carried out separately for this sample of elementary, middle, and high schools, while not quantitatively representative of public schools in the nation, can suggest possible generalizations to other American schools. In any case, separate analyses by grade level are essential in using the SASS student achievement subfile, not only because different factors are related to achievement at different levels, but also because different achievement measures are used in each state at the different school levels

Finally, because the sample consists of a nonrandom subset of 20 states, no claims can be made that estimates of effect-sizes are quantitatively representative of the nation. The states included in the file are Alabama, California, Delaware, Florida, Georgia, Hawaii, Kentucky, Louisiana, Maine, Massachusetts, Michigan, Montana, New Hampshire, New York, Pennsylvania, Rhode Island, Tennessee, Texas, Utah, and Washington.

A School-level Measure of Student Achievement

The first step in rendering state assessment scores comparable is to compute each school's score as it relates to other schools' scores in the state at the same grade. That is, the (unweighted) mean score of the schools in the database for the particular grade and state is subtracted from each of the scores to create a score with a mean of 0 at each state and grade; then these scores are divided by the standard deviation of the school scores to create a score with a standard deviation of 1.0.

Using this measure, third grade reading scores in one state, fourth grade reading scores in another, and fifth grade reading scores in a third are taken to be comparable achievement measures for the purpose of computing within-state correlations across elementary schools with factors such as average class

---

[6] The SASS student achievement subfile included 347 ungraded schools, or 14 percent of the total, a much larger representation of ungraded schools than in the population (3 percent). However, ungraded schools were not included in the analyses reported in this study.

[7] Additional descriptive comparisons of the schools in the SASS student achievement subfile with other public schools are reported by Wu, Royal, and McLaughlin (1997).

size and school behavioral climate. All achievement differences between states are removed in this measure, so comparisons with school characteristics would also need to remove between-state variation in school characteristics.

The second step is, therefore, to re-introduce between-state variation using a common standard, State NAEP. In a separate study, the State NAEP schools were linked to state assessment scores; and the means, standard deviations, and correlations of State NAEP school means with state assessment school means were computed. Those results were used in this study to create reading and mathematics achievement measures that include (1) between-state variation in means, (2) between-school variation proportional to between-school variation among State NAEP schools, and (3) a factor that attenuates within-state variation for states in which the assessment is only moderately correlated with State NAEP. The effect of the third factor, multiplying by the correlation between the NAEP and state reading assessment scores, "projects" the state assessment variation onto the NAEP scale, capturing that part of the state assessment score that is like NAEP.[8]

Thus, this achievement score "spreads schools apart" in states in which (a) school NAEP scores are more varied and (b) the state assessment appears to be measuring skills highly related to NAEP. The effect of this spreading is to give greater weight to variations within these states in the estimation of correlations of achievement with SASS measures.

Important assumptions are needed to apply the NAEP adjustment to the scores at grades other than fourth and to mathematics scores. The between-state NAEP adjustment was based on the 1994 State NAEP grade 4 assessment for reading and on the 1992 and 1996 State NAEP grades 4 and 8 assessments for mathematics. Application of these adjustments to state assessment scores in middle school (for reading) and high school is based on the assumption that variation in achievement *between states* is stable across grades. The 1992 and 1996 State NAEP mathematics assessment results support this assumption, in that the correlations between the grade 4 and grade 8 state means are 0.95 and 0.92, respectively (Mullis, Campbell, and Farstrup 1993; Reese et al. 1997).

---

[8] An alternative to projecting the scores onto the NAEP reading scale would be to omit the third factor, treating each state's reading assessment as the relevant outcome for that state. However, it would be more difficult to characterize an achievement measure that is based on different scales across states.

However, no information is available regarding State NAEP means at the high school level.

The use of 1992 and 1996 State NAEP state means and standard deviations to construct the 1994 mathematics adjustment assumes that state means varied smoothly, if at all, from 1992 to 1994 to 1996. In fact, the correlations of this study's 20 State NAEP mathematics means between 1992 and 1996 were 0.91 for grade 4 and 0.93 for grade 8, suggesting that interpolating 1994 figures from 1992 and 1996 figures is reasonable.

Finally, the correlations used in the adjustments were based on grade 4 reading assessments. Use of these correlations in the adjustment of within-state variation in math scores assumes that between-state variations in reading and math are highly correlated. Because state assessments usually combine reading and math tests from the same publisher and in the same testing session, it is plausible to assume that factors that would affect the reading correlations in different states (e.g., the reliability of the state assessment instrument and distribution of state assessment scores) would also affect the math correlations. The 1992 State NAEP assessments in reading and mathematics in grade 4 indirectly support this assumption, in that the correlation between reading and mathematics state means is 0.94 (Mullis et al. 1993). Nevertheless, the question remains whether the results of substantive analyses will be diminished by the extrapolation of between-state variation in average achievement from grade 4 to middle and high school variation. Comparative analyses of NAEP-adjusted vs. pooled within-state findings across school levels (carried out in this study) address this question.

Although State NAEP data were used to capture between-state variation in achievement, it would be highly misleading to interpret the SASS student achievement measures as a surrogate of the school's average NAEP proficiency. First, State NAEP differs from individual state assessments in student sampling (each student takes only a fraction of the NAEP test), administration (a federal government contractor trains test administrators and monitors many testing sessions), motivation (NAEP is a low-stakes assessment with no individual student or school reporting), and item formats (NAEP has a substantial portion of extended open-ended items). The achievement measure developed for this study yields an unbiased estimate of school NAEP means, but with different standard errors in each state. The measure is not dependent on evidence that NAEP is *equated* to the various state assess-

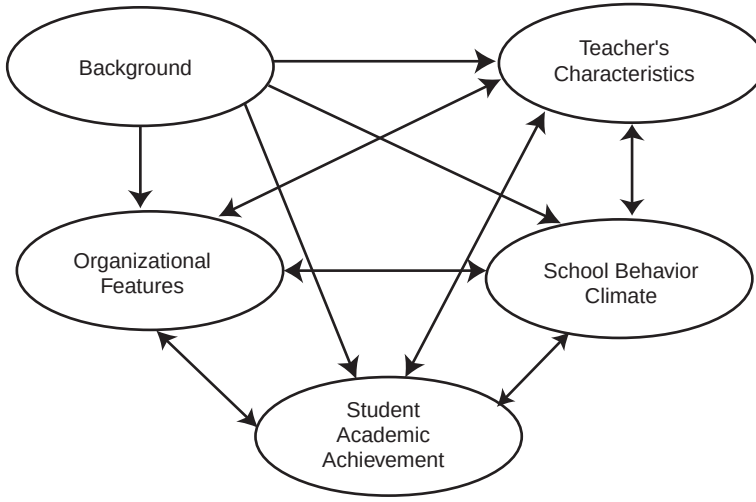
ments, and evidence that they might be "equatable" (i.e., parallel) was not sought. In fact, it is highly unlikely that the states' individual assessments are parallel to NAEP, due to differences in administration, item format, and content frameworks. In other research, we have shown that it is feasible to project state assessment results onto the NAEP scale without assuming that the tests are parallel (McLaughlin 1998). Second, the correlations between NAEP and state assessments differ substantially between states. Although the median correlation in these states in 1994 was 0.70, the smallest three correlations were between 0.30 and 0.50 (Wu, Royal, and McLaughlin 1997). Within-state variation of these synthetic NAEP school means will be smaller than variation of actual NAEP school means, especially in states where assessments are not highly correlated with NAEP.

School-level Correlates of Achievement

Student academic performance is shaped by multiple factors relating to the school, teaching process, students' social and family backgrounds, and community; also, a school's reputation for academic performance can affect parents' decisions, students' behavior, and teachers' attitudes and decisions. We model student achievement in American public schools as related to four types of factors: (a) students' background, (b) four organizational features of the school, (c) professional characteristics of the teachers, and (d) school behavior climate. While all these factors affect student academic success, they also interact with each other; and organizational characteristics and teacher choices can be affected by achievement at the school. We therefore conceptualize the interrelationships among the five categories in this model as a web of interactions. Figure 1 graphically describes the general model.

The model shown in figure 1 refers to the school as a unit. Of course, achievement is an individual student characteristic, and the majority of variation in achievement is among students in the same classroom and between classrooms in the same school. Nevertheless, there is substantial reliable variation in achievement between schools; and from a policy perspective, there may be actions that can improve the overall achievement level in a school. While analyses at the school level may not shed light on individual variation in learning, they can provide evidence on the correlation between school reform policies and achievement. Hierarchically structured data, with both individual student data and schools and staffing data, such as collected in NAEP and NELS:88, facilitate understanding of the correlates of individual student achievement;

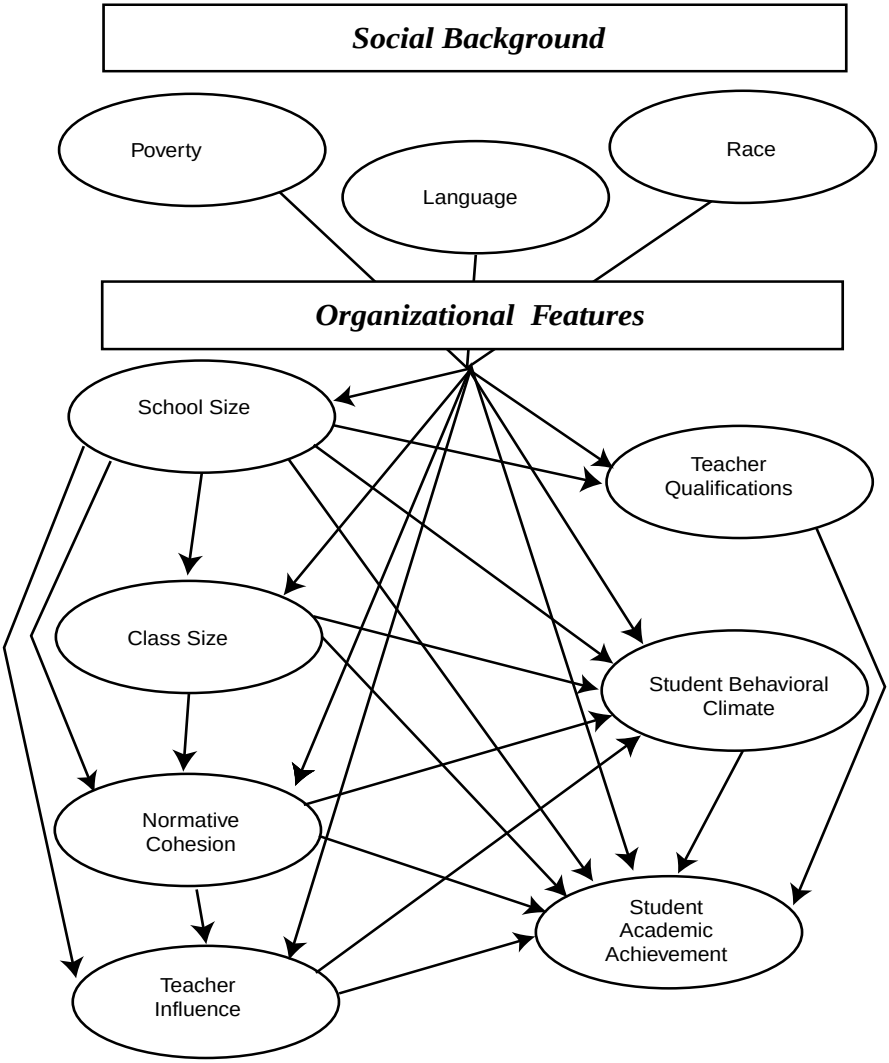
Figure 1. School-level Correlates of Student Achievement: General Model



but these data are much more costly to collect on a school sample the size of SASS than is the construction of a synthetic achievement measure from existing NAEP and state assessment data. In any case, the existence of within-school variability does not threaten the validity of analyses at the school level.

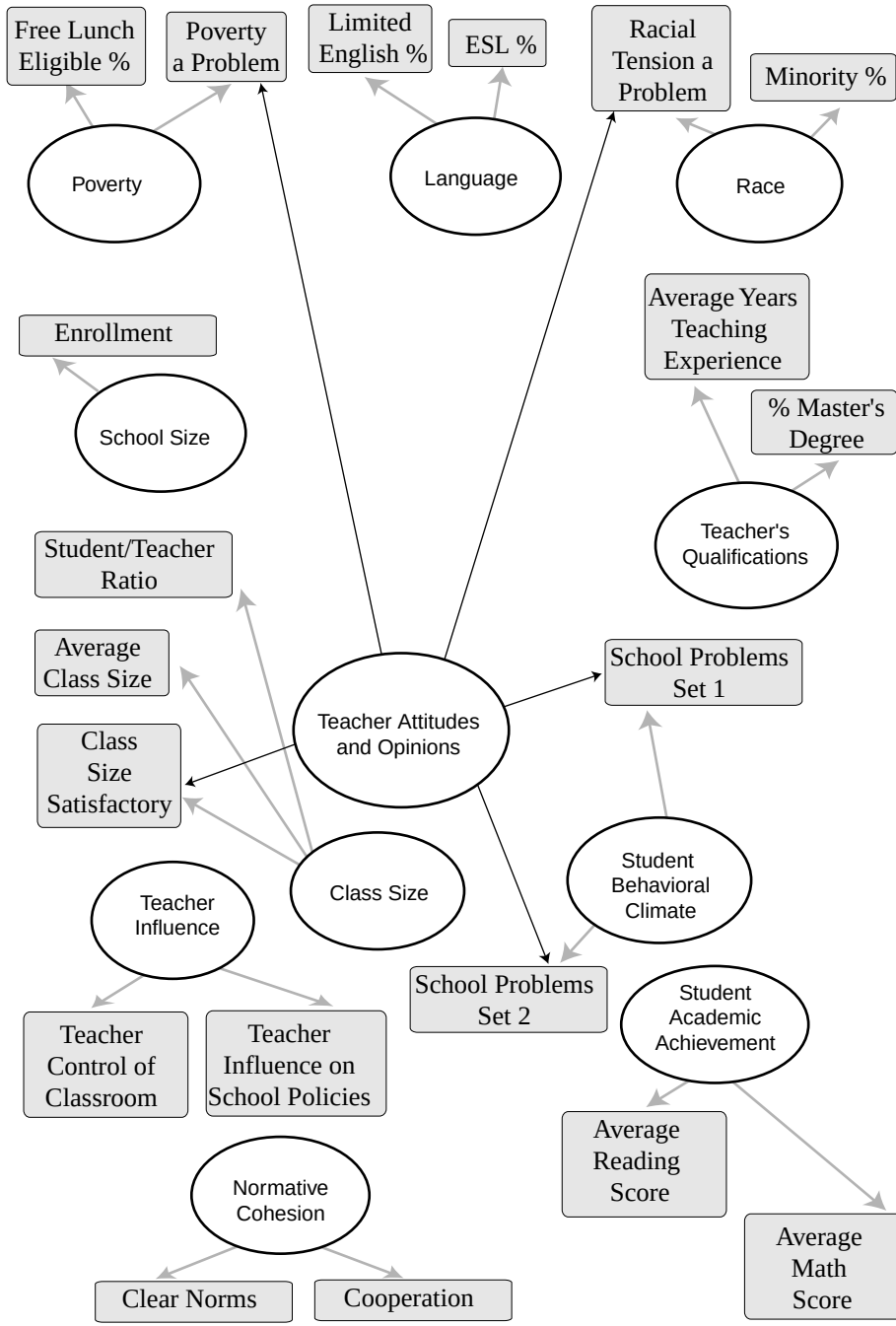
Using the SASS student achievement subfile, the model in figure 1 can be further specified in a variety of ways, one of which is shown in figure 2. The background category in the model is represented by three factors: (1) percentages of students in poverty, (2) percentages with language barriers, and (3) percentages in racial-ethnic minorities. The organizational category is represented by four factors: (1) school size (total enrollment), (2) average class size, (3) teachers' perceived influence, and (4) normative cohesion. The aim of the analyses to be carried out is to test hypotheses about the relations among these factors, either in terms of partial correlations or in terms of fits of linear models, such as that represented graphically in figure 2. By testing these models, inferences can be made about correlates of achievement across a wide range of public schools, although the "arrows" cannot, in most cases, be taken to indicate a direction of causality because of the alternative explanations of many of the correlations.

Figure 2. One Possible Path Model Relating Achievement and School and Background Factors



Each of the factors in the model, except school size, is represented by multiple measures in SASS, as indicated in figure 3. In structural equation terminology, figure 3 presents the measurement model corresponding to the structural model in figure 2. The arrows in figure 3 indicate assumptions about the sources of variance in the observed measures. Each of the factors in figure 2, except school size, is represented by at least two indicators, providing the

Figure 3. Measurement Model for School-level Correlates of Achievement



capacity for estimating the contribution of measurement error to variance in the indicators. Also indicated in figure 3 is a factor, *teachers' attitudes and opinions*, representing a common response pattern among five of the measures. These measures may be more positively intercorrelated than other measures because they all represent teachers' subjective opinions about their schools.

The two indicators of school behavior climate are based on 20 items concerning teachers' perceptions of problems in the school. The two parallel measures were constructed by averaging balanced halves of these items. For example, drug abuse is in one set, alcohol abuse in the other; student absenteeism in one and dropping out in the other; vandalism in one and robbery in the other. Two topics for which there were multiple items, tardiness and attacks on teachers, were included in both sets.

Determining the correlates separately for elementary, middle, and high schools provides an opportunity to explore the patterns of change in the correlations over the school years. Much like Herriot and Firestone's (1984) arguments that the images of schools vary by level[9] and that each school level operates differently,[10] we anticipate that the relative importance of the various factors to student academic achievement will vary by school level. For example, we anticipate that the relationship between social background factors and student performance will be found to be greater in elementary schools than in secondary schools.

Student Background

Family background and socioeconomic status are consistently shown to be related to student achievement (Coleman et al. 1966; Hanushek 1986). While the major purpose of research on schools is to identify characteristics of *schools* that contribute to student achievement, it is essential to take background characteristics into account because they affect the intercorrelations of school measures. For example, suppose that poor children were found to be attending schools with chipping paint. We would expect to find a correla-

---

[9] Elementary schools are imagined to be more rational and bureaucratic, while high schools are seen as more anarchic and envisioned to be a loosely coupled system.

[10] In elementary schools, curriculum is more limited, while in high schools, curriculum is broad. Also, the operations of elementary schools are more centralized and consensus-driven, while those in high schools have high levels of complexity and differentiation.

tion between chipping paint and test scores, but one should not infer from that correlation that painting schools will improve test scores. It is the pattern of correlates among schools with students with similar backgrounds that is of interest. One need not address social issues about the sources of the impact of background variations on achievement in school to realize the need to control for these factors in modeling school processes.

It is particularly important to include background factors in this study of school-level factors, because the database does not contain longitudinal achievement data on individual students or student cohorts. School characteristics and policies are more likely to be correlated with school-level variation in gains in achievement than with achievement differences measured at one point in a student's career. Including student background measures in the model serves to control for much of the between-school variation in achievement potential that students bring to school. Nevertheless, the lack of "pre" measures of achievement underlines the need to interpret the results of analyses of SASS data relevant to the model in figures 1 and 2 as causally indeterminate.

Differences in the family backgrounds of students in different schools are reflected in three factors: (1) poverty, (2) English language proficiency, and (3) racial-ethnic minority status. The level of *poverty* in a school is measured by two indicators available from the SASS database: (1) the percentage of students who qualify for the national school lunch program and (2) teachers' identification of poverty as a problem. The level of *English language proficiency* in the school is composed of two similar indicators: (1) the percentage of students identified as Limited English Proficient (LEP) and (2) the percentage of students participating in the school's English-as-a-Second-Language (ESL) program. Finally, the *minority status* of a school is measured by (1) the percentage of white students in the school and (2) teachers' identification of racial tension as a problem.

A unique strength of the SASS database is that background factors can be measured by an objective indicator, the percentage of students with the corresponding characteristic, and by the perception of a sample of teachers in the school that said factor is a source of problems in the school. For example, the underlying factor of poverty is only imperfectly measured by the percentage of children who are eligible for the federal free lunch program, due to differences in cost of living and in eligibility counting procedures between school districts. Incorporating the perceptions of a sample of teachers in the

school that poverty is a problem in the school can eliminate some of the error of measurement of poverty. Of course, there are different kinds of error in teacher perceptions, ranging from different interpretations of survey items to sampling error, but the combination of the two indicators can be expected to have greater validity for the impact of poverty on learning in the school than either separately. Because SASS has many related objective and subjective measurements, one can control for measurement error associated with variation in teachers' use of the response scale (e.g., some teachers mean something more serious by "problem" than others do) by estimating the extent to which each teacher tends to be a positive or negative responder to attitude and opinion items.

Alternative indicators of a particular factor, such as subjective and objective assessments of poverty, must be correlated, but they need not be highly intercorrelated; however, a low intercorrelation is likely to limit the power of the data to measure correlations with achievement. The intercorrelations of the indicators included in each composite factor are shown in appendix A following this paper. For example, for poverty, the intercorrelations between the objective and subjective indicators are 0.55, 0.51, and 0.47 for elementary, middle, and secondary schools, respectively.[11] These intercorrelations are themselves limited by the "reliability" of the indicators that are based on averages of teachers' responses. If different teachers see the same school very differently, the average of their responses cannot tell a great deal about the school, per se. The percentage of variance in school means that is attributable to between-school variation, which is a measure of this reliability, is also shown in appendix A. For the average rating of whether poverty is a moderate or serious problem, the reliabilities are 0.73, 0.79, and 0.84 for teachers in elementary, middle, and high schools, respectively.

School Organizational Features

Both objective features of a school's organization such as its class sizes and subjective features such as the level of cooperation among its staff may be correlated with achievement. However, unlike the social background factors,

[11] The corresponding intercorrelations for the minority and language factors are approximately 0.45 and 0.75. The language factor, unlike the other background factors used here, is defined by two objective indicators.

these features are endogenous and, to a varying extent, under control of the principals and teachers in the school. For example, a magnet school with a reputation for attracting students with potential for careers in science is likely to have higher test scores than other schools, purely as a function of the backgrounds of the students who enroll; and its magnet status may affect class sizes, either enlarging them to respond to demand or lowering them as a result of special funding as a magnet school. Although the specific models presented in this report focus on accounting for variation in achievement in terms of variation in organizational characteristics, the direction of causality is not determined in these data.

SASS has a wide range of information about schools obtained from the principals in the administrator and school questionnaire and from the teachers in the teacher questionnaire. Four organizational features have been selected for inclusion in the model for this report: (1) school size, (2) average class size, (3) teachers' sense of their influence over school affairs, and (4) normative cohesion of the school's staff. A variety of other SASS organizational measures, such as organizational complexity as reflected in the number of different kinds of positions in the school, organizational goals as expressed by the principal, perceptions of outside influence on decisions by state agencies and local school boards, and staff diversity, might be included in a more elaborate model.

School size, as measured by total enrollment, has been shown to have a significant effect on the school's performance, yet the direction of the effect has not been consistent. Because school size has different implications for instructional resources at the elementary and secondary levels, and because school size is highly correlated with the sizes of classes in the school, the correlations of school size and achievement measures vary with the type of schools studied and the variables included in the study.

Class size, according to common sense, should have an effect on students' academic performance. However, it has proven difficult to isolate and demonstrate that effect. SASS has three indicators of a school's average class size: (1) average class size as reported by the sampled teachers;[12] (2) the school's

[12] For a teacher teaching a single self-contained classroom, class size is the total number of students enrolled in the teacher's class at the school. By contrast, for a teacher teaching multiple departmentalized classrooms, class size is calculated as the average number of students in the teacher's classes.

teacher/student ratio, calculated as the total number of students in school divided by the sum of the full-time teachers in the school and one-half the number of part-time teachers in the school; and (3) the average of the sampled teachers' ratings of satisfaction with the size of their classes.

As a measure of the extent to which the classrooms in a school have too many students for optimal learning, each of these has a different source of measurement error, such as variation in staff counting methods, teacher sampling variation, invisibility of class size remedies (e.g., teacher aides and parent volunteers), and different criteria for teachers' satisfaction. Although none of these by itself is a perfect measure of a school's class sizes, their combination may provide a more valid measure of the impact of class size on student learning than any of them considered separately.[13]

Teachers' sense of influence, measured by the average of sampled teachers' responses to SASS questions about their perceptions that they have influence over school policies[14] and over matters concerning their own class[15], differentiates schools with varying management styles and teacher roles. Although this factor may not have a measurable direct effect on achievement, the sense of efficacy represented by this factor may be related to the general climate in the classroom, which in turn can affect learning.

Finally, *normative cohesion* of the staff refers to the cultural solidarity among staff members in the school or the collective norms that govern staff behavior in this organization and may also be correlated with the climate in classrooms in the school. Normative cohesion can be measured by two SASS

---

[13] Among elementary schools, the reliability estimate of teacher-reported class sizes as an indicator of the school is low, 0.31, and the intercorrelations of that with the other class size indicators are also low, 0.28 and 0.29. This may limit the potential elementary school correlations of class size with achievement.

[14] Included items refer to influence over school discipline policies, the content of inservice programs, hiring new full-time teachers, deciding priorities in spending the school budget, evaluating teachers, and establishing a curriculum. The reliability of the school mean is about 0.62.

[15] Included items refer to control in one's own classroom over selection of textbooks and other instructional materials, selection of contents, topics, and skills to be taught, selection of teaching techniques, evaluating and grading students, disciplining students, and determining the amount of homework to be assigned. The reliability of the school mean ranges from 0.36 at the elementary level to 0.50 at the secondary level.

composite measures: (1) a score for the clarity of norms[16] and (2) a score for staff cooperation.[17]

Teachers' Qualifications

Common sense leads to the expectation that more qualified teachers create more effective learning environments in their classrooms, so part of the variation in achievement between schools may be due to teachers' qualifications. In SASS, a variety of teachers' characteristics are recorded for samples of five teachers per school, on average. Because level of education and amount of teaching experience are widely used to determine the pay scales of teachers, these form a logical basis for measuring teacher qualifications. In particular, the school level measures are (1) average years of teaching experience and (2) the percentage of teachers who acquired at least a master's degree.

Although teachers' qualifications are included in the analysis, the factor is relatively weak, compared to the other factors.[18] A wide variety of other SASS measures of teacher qualifications should be included in a study focusing particularly on teaching quality and achievement, including out-of-field teaching, selectivity of the colleges the teachers attended, amount of inservice training, hours spent on school-related activities, and training to teach Limited English Proficient (LEP) students.

---

[16] Calculated as the average of teachers' responses to SASS questions about sharing beliefs and values with colleagues in school, receiving support from parents, goals and priorities in school being clear, rules being consistently enforced by all teachers, principal backing up staff members, principal letting staff know what is expected of them, and principal having a clear vision of the type of school wanted and communicating this model to staff. The mean correlation among these seven measures is 0.42, and the reliability of the school mean is about 0.57.

[17] Calculated as the average teacher's responses to SASS questions about getting cooperative effort among staff members, making a conscious effort to coordinate course content with other teachers, planning with media specialist or librarian an integration of their specialty into teaching, and viewing the behavior of school administration as being supportive and encouraging. The mean correlation among these four measures is 0.20, and the reliability of the school mean is about 0.48.

[18] There is substantial within-school variation in years of teaching experience (reliabilities of 0.29, 0.30, and 0.37 at the three grade levels); correlations with the percent of responding teachers with a master's degree range from 0.14 to 0.27.

The relation between a school's averages of teachers' qualifications and achievement is complex. It may reflect choices by teachers of where to teach and of school districts as to how to allocate resources, as well as the impact of experience and training on effective learning environments. As with school organizational factors, the direction of causality of the relations between teacher qualifications and achievement cannot be inferred directly from correlations, but these correlations provide valuable evidence of relations that must be explained in some manner.

School Climate

Student achievement is difficult when the climate in a school reflects factors such as drugs, violence, vandalism, truancy, lack of respect for teachers, and lack of enthusiasm for learning. These characteristics are difficult to measure in a uniform manner across a large sample of schools, but SASS has attempted such a measurement by asking teachers to indicate which of two dozen different types of potential problems are serious, moderate, minor, or nonexistent in their schools. To assess the extent of measurement error in these perceptions, two composite measures can be constructed by arbitrarily dividing the problem ratings into halves.[19]

Although a positive relationship between school climate and achievement can be expected, the direction of causality in this relation is particularly ambiguous. There is likely to be a positive feedback between students' focus on achievement and teachers' perceptions of their behaviors and demeanors. Nevertheless, evidence about the significance of this correlation, and of its mediating role in other relations with achievement in schools, is valuable.

Analytical Method

The SASS student achievement database contains school-level statistics on hundreds of measures for over 2,000 public schools in 20 states. The potential for analyses of this database is enormous. In this report, we have constructed a set of 18 composites of SASS items; and for a baseline demonstration of analytical feasibility, the database has essentially been reduced to the intercorrelation matrix of these 18 composites, along with two assessment

[19] The reliabilities of the two climate composites range from 0.70 to 0.86 for the three school levels, and their intercorrelation ranges from 0.50 to 0.72.

scores, for schools at each of three grade levels, plus, for the purpose of standard error estimation, an indicator of the state in which each school resides.[20] However, the raw correlations between the 18 composites and the two achievement measures do not provide meaningfully interpretable information, because many of the apparent correlations are mere reflections of correlations among other measures, and other "real" correlations are masked by confounding measures and can only be uncovered by controlling for those confounders. Thus, the first meaningful stage of analysis is to examine partial correlations between achievement measures and SASS composite measures of school organization and climate and teacher qualifications, controlling for social background factors.

Partial correlation analysis, as a method of testing for significant relations between school composites and achievement measures, has the advantage that it is neutral with respect to causal ordering; but the picture of the correlates of achievement provided by partial correlations is limited, in that multivariate structure is not apparent. For example, the partial correlations of both normative cohesion and school climate with achievement may be positive, but one cannot discern from them whether this is due to a common factor in climate and cohesion or to independent factors. A form of multivariate regression is needed to identify the more complex structure.

The simplest form of multivariate regression is ordinary least squares (OLS) linear regression. This methodology can reveal the multivariate structure when its assumptions are satisfied, but an important assumption in OLS modeling that limits its value for educational research is that the "predictor" measures are measured without error. With many databases, this assumption is untestable, because only a single measure of each construct is available. However, the SASS database with its multiple sources of information about school-related constructs offers the opportunity to take measurement error into account when modeling the structural relations among factors. Structural equation modeling (SEM) jointly models the structural relations among latent factors (as in figure 2), while simultaneously modeling measurement error (as in fig-

[20] Although exploration of other functional forms may yield additional insights, estimation purely in terms of linear models is an efficient initial step, because software packages are readily available and most important relations among educational factors are monotonic and therefore "visible" to linear analyses.

ure 3). Computer programs such as LISREL, EQS, and SAS PROC CALIS can be used to estimate the variance components in SEM (see Bollen and Bollen 1989). The primary analytical results presented in this report are SEM analyses based on SAS PROC CALIS estimation.

SEM is particularly helpful in specifying, estimating, and testing hypothesized relationships among meaningful concepts, or factors,[21] by allowing such concepts to be estimated from several indicators, or measures. In SEM, the variance of a latent variable reflects the variation on the common factor among indicators of that latent variable, as measured by their intercorrelations. In SASS, the indicators themselves can be determined as composites of responses to individual items. For this study the construction of indicators was based on both judgment and factor analyses of items.

The structural model specifications for estimation of student achievement correlates corresponding to figure 2 are given by the following set of equations. A complementary measurement model relates each of the latent predictors to multiple SASS measures with the exceptions that reading achievement, mathematics achievement, and school size are each based on a single measure.

$$\eta_{\text{Reading}} = \Gamma_{11}\eta_{\text{School climate}} + \Gamma_{12}\eta_{\text{Tchr qualifications}} + \Gamma_{13}\eta_{\text{Teacher control}} + \Gamma_{14}\eta_{\text{School size}} + \Gamma_{15}\eta_{\text{Class size}} + \Gamma_{16}\eta_{\text{Norm cohesion}} + \Gamma_{17}\xi_{\text{Language}} + \Gamma_{18}\xi_{\text{Poverty}} + \Gamma_{19}\xi_{\text{Race}} + \delta_1$$

$$\eta_{\text{Mathematics}} = \Gamma_{21}\eta_{\text{School climate}} + \Gamma_{22}\eta_{\text{Tchr qualifications}} + \Gamma_{23}\eta_{\text{Teacher control}} + \Gamma_{24}\eta_{\text{School size}} + \Gamma_{25}\eta_{\text{Class size}} + \Gamma_{26}\eta_{\text{Norm cohesion}} + \Gamma_{27}\xi_{\text{Language}} + \Gamma_{28}\xi_{\text{Poverty}} + \Gamma_{29}\xi_{\text{Race}} + \delta_2$$

$$\eta_{\text{School climate}} = \Gamma_{31}\eta_{\text{Tchr control}} + \Gamma_{32}\eta_{\text{School size}} + \Gamma_{33}\eta_{\text{Class size}} + \Gamma_{34}\eta_{\text{Norm cohesion}} + \Gamma_{35}\xi_{\text{Language}} + \Gamma_{36}\xi_{\text{Poverty}} + \Gamma_{37}\xi_{\text{Race}} + \delta_3$$

$$\eta_{\text{Tchr control}} = \Gamma_{41}\eta_{\text{School size}} + \Gamma_{42}\eta_{\text{Norm cohesion}} + \Gamma_{43}\xi_{\text{Language}} + \Gamma_{44}\xi_{\text{Poverty}} + \Gamma_{45}\xi_{\text{Race}} + \delta_4$$

$$\eta_{\text{Tchr qualifications}} = \Gamma_{51}\eta_{\text{School size}} + \Gamma_{52}\xi_{\text{Language}} + \Gamma_{53}\xi_{\text{Poverty}} + \Gamma_{54}\xi_{\text{Race}} + \delta_5$$

$$\eta_{\text{Norm cohesion}} = \Gamma_{61}\eta_{\text{School size}} + \Gamma_{62}\xi_{\text{Language}} + \Gamma_{63}\xi_{\text{Poverty}} + \Gamma_{64}\xi_{\text{Race}} + \delta_6$$

$$\eta_{\text{Class size}} = \Gamma_{71}\eta_{\text{School size}} + \Gamma_{72}\xi_{\text{Language}} + \Gamma_{73}\xi_{\text{Poverty}} + \Gamma_{74}\xi_{\text{Race}} + \delta_7$$

$$\eta_{\text{School size}} = \Gamma_{81}\xi_{\text{Language}} + \Gamma_{82}\xi_{\text{Poverty}} + \Gamma_{83}\xi_{\text{Race}} + \delta_8$$

[21] Technically referred to as latent variables, yet also known as unobserved or unmeasured variables.

Because the data are correlational, it should be pointed out that the analysis is also consistent with views that characteristics like school climate, cohesive norms, cooperation, and satisfaction are affected by the kinds of skills and attitudes that children bring to the school, which are best reflected in achievement scores. For example, if the student peer norm is to focus on class work, there is likely to be less of a problem with absenteeism, tardiness, and class cuts. Likewise, a negative correlation between average class size and average achievement may indicate that smaller classes facilitate achievement, but it may also be due to a socioeconomic variation between school communities that affects both class size and achievement. Nevertheless, the methodology for the analyses is a variant of linear modeling with asymmetric "independent" and "dependent" variables. Therefore, although the correlational results may appear to be couched in terms of "effects" of some factors on others, these "effects" merely indicate the partitioning of the variance in the "dependent" factors into covariances with "independent" factors. This use of the term "effect" should not be confused with its use in the description of causal relations. Without carefully controlled experimental studies, the direction of causality cannot be inferred, only conjectured.

In addition to SEM analyses and in order to verify the robustness of the findings, we also carried out ordinary least squares (OLS) linear regression analyses.[22] For these analyses, we created composites for poverty, English language proficiency, racial-ethnic minority, teacher qualifications, school climate, school size, class size, teacher control, and normative cohesion, using the same measures as in the SEM analyses. Separate regressions were performed for elementary, middle, and secondary schools.

---

[22] The major difference between the two methodologies is in how they treat measurement error in predictive measures. The basic assumption of OLS linear regression is that predictors are measured "without error." That is, each estimated coefficient represents the extent to which a measure *per se*, not an underlying construct that it measures, accounts for variation in the dependent variable. SEM analysis, on the other hand, accounts for measurement error and allows for the specification of correlations among both the constructs and the measurement errors. On the other hand, SEM's greater flexibility must be weighed not only against the additional computational complexity, but also against the additional complexity of interpretation. Relations among latent variables do not have the same simplicity as relations among observable variables. Therefore, our approach is to examine and compare the results of both methodologies to identify robust patterns in the correlates of school-level achievement.

Finally, because this report was not focusing on differences in achievement correlates between states, the powerful analytical technique of hierarchical linear modeling (HLM) was not used. Hierarchical modeling would be a powerful tool in the use of these data to study the effects of state educational policies on school factors and achievement.

The value of the SASS student achievement subfile depends on whether meaningful patterns of statistically significant results emerge from analyses of relations between SASS measures and achievement. Thus, assessment of statistical significance is central to the study. Unfortunately, the usual tests of statistical significance available in common statistical packages are based on the assumptions of simple random sampling, and the SASS student achievement subfile is far from a simple random sample. Because of the similarity of the unweighted school sample to the universe of American public schools, unweighted analyses are appropriate.[23] However, the variance components within and between states cannot be expected to be uniform. Therefore, to provide valid estimates of the standard errors of statistical estimates, for the purpose of statistical significance testing, another method is needed. For this study, standard errors for all statistics were estimated by repeating each analysis on 100 random half-samples of schools. The standard deviation of the statistics computed for the various half-samples provided valid estimates of the corresponding standard errors. Because there is systematic variance in achievement measures between states (i.e., there is significant variation among NAEP mean state scores), it was necessary to select the random half-samples by state. That is, each half-sample consists of all the schools in the database for a randomly selected half of the 20 states. Although this method of standard error estimation is not as computationally efficient as balanced repeated replications, it provides valid standard error estimates.

Results

The major question addressed by these analyses of the SASS student achievement subfile is: Are organizational factors, teachers' qualifications, and

[23] To support reporting state-by-state statistics, SASS purposely (proportionately) oversamples schools in less populous states. As a result, differential weights are needed to estimate descriptive statistics for groups of states. Because the purpose of this study is not to produce descriptive statistics and because differential weighting substantially reduces the precision of estimates, differential weights are not used in this study.

school behavioral climate correlates of school mean assessment scores? The value of the SASS student achievement subfile is tested by the answers it gives to this question. Although the analyses reported here merely scratch the surface of the potential for analyses of these data, they should provide evidence of a meaningful pattern of relations between school-level factors and assessment scores.

The starting point for this research is the assumption that there is meaningful variation in assessment scores between schools. The choice of educational policies depends on the extent to which that variation is attributable to factors that are under a local school system's control, as compared to factors associated with the communities in which the schools are located. Estimation of the relative correlations of background and school-based factors with achievement is not straightforward, because background factors at least partially determine the levels of school-level factors. If, for example, teacher qualifications are correlated with student achievement, this may be due both to the fact that more qualified teachers teach more effectively and to the fact that higher achieving schools can attract more qualified teachers. Nevertheless, any analyses of school-based factors must control for background differences.

Three categories of school-based factors are included in these analyses: (1) school behavior climate, based on teachers' perceptions of problems in the school, (2) teachers' qualifications (that is, their years of teaching experience and attainment of a master's degree), and (3) four organizational characteristics—school and class sizes, normative cohesion, and teachers' sense of control and influence. As a first step in exploring the relations between these factors and achievement scores, partial correlations of the school-based SASS factors[24] with reading and math assessment scores and with each other are shown in table 1. These partial correlations represent the bivariate relations among the factors, partialing out the three background factors of poverty, language barriers, and race-ethnicity.

Partial correlations reveal the contributions of the school-based factors considered singly to achievement variance. However, they do not capture multivariate relations among the school-based factors. For example, are the negative partial correlations of school and class size with reading assessment scores in

[24] Each factor is defined as the average of the measures indicating the factor shown in figure 3, with measures scaled to equal standard deviations.

Table 1. Partial Correlations of Mean Reading and Mathematics Scores with School-Level Factors in Public Elementary, Middle, and Secondary Schools, Controlling for Background Factors

Factor 1	Factor 2	Elementary (n = 1123)		Middle (n = 496)		Secondary (n = 595)	
		Reading	*Math*	*Reading*	*Math*	*Reading*	*Math*
	Student Achievement in Reading & Mathematics						
School Size		–0.10	–0.09	–0.18**	–0.17	+0.11	–0.06
Class Size		–0.10	–0.02	–0.26*	–0.10	–0.12*	–0.11**
Normative Cohesion		–0.01	+0.01	+0.04	–0.06	–0.05	–0.09
Teachers' Influence		+0.03	+0.05	+0.09	+0.20*	+0.07	+0.23*
Teachers' Influence		–0.02	–0.04	–0.05	–0.10	–0.08	–0.16*
Teachers' Qualifications		+0.04	+0.01	+0.12*	+0.05	+0.08*	+0.11*
	School Climate						
School Size		–0.14*		–0.26*		–0.36*	
Class Size		–0.15		–0.24*		–0.26*	
Normative Cohesion		+0.34*		+0.41*		+0.44*	
Teachers' Influence		+0.18*		+0.20*		+0.26*	
Teachers' Qualifications		+0.00		+0.06		–0.08	
	Teachers' Self-Perceptions of Influence						
School Size		–0.13*		–0.14*		–0.22*	
Class Size		–0.09		–0.03		–0.14*	
Normative Cohesion		+0.35*		+0.27*		+0.25*	
Teachers' Qualifications		–0.10*		–0.06*		–0.06	
	Normative Cohesion						
School Size		+0.02		–0.01		–0.09*	
Class Size		–0.07		–0.03		–0.07	
Teachers' Qualifications		–0.01		–0.04		–0.03	

Table 1. Partial Correlations of Mean Reading and Mathematics Scores with School level Factors in Public Elementary, Middle, and Secondary Schools, Controlling for Background Factors (continued)

Factor 1	Factor 2	Elementary (n = 1123)	Middle (n = 496)	Secondary (n = 595)
	Teachers' Qualifications			
School Size		+0.11	+0.10	+0.24
Class Size		+0.06	–0.04	+0.11
	Class Size			
School Size		+0.33*	+0.48*	+0.54*

NOTES: Table entries are partial correlations, partialing out poverty, language, and race-ethnicity factors.

(*) p<.05 based on repeated half-sample standard deviations.

middle schools independent effects, or is one of the correlations a byproduct of the high intercorrelation between school and class size? Multiple regression and structural equation modeling (SEM) provide a more interpretable picture of the interrelations among the factors.

Ordinary least squares (OLS) multiple regression is a second step in the analysis of school-based correlates of achievement. Separate equations can be modeled for reading and mathematics assessment scores and also for school climate, normative cohesion, teachers' influence, class size, and teachers' qualifications. Unlike partial correlations, multiple regression differentiates between "predictor" and "dependent" variables, although the interpretation of equations remains merely that a combination of predictors is correlated with the dependent variable. Estimates of standardized regression coefficients for the equations indicated in figure 2 are shown in table 2.[25]

An important limitation of OLS regression is that while variance in each (intermediate) endogenous factor is partially accounted for by other factors, each factor is assumed to be measured without error in predicting other factors. For survey measures, that assumption is not supportable; and a preferable method of analysis is structural equation modeling (SEM), which takes measurement error into account in estimating the proportions of variance in factors

[25] The same factor definitions are used in tables 1 and 2. Estimates of coefficients for background factors are not shown in table 3 because the coefficients are not germane to the exploration of school-based correlates of achievement

Table 2. OLS Standardized Regression Weights for School-level Factors Associated with Mean Reading and Mathematics Scores in Public Elementary, Middle, and Secondary Schools

Factor 1	Factor 2	Elementary (n = 1123)		Middle (n = 496)		Secondary (n = 595)	
		Reading	*Math*	*Reading*	*Math*	*Reading*	*Math*
	Achievement in Reading & Student Mathematics						
School Size		-–0.0	–0.07	–0.04	–0.08	+0.29*	+0.11
Class Size		–0.05	+0.01	–0.16*	–0.04	–0.20*	–0.08*
Normative Cohesion		–0.02	+0.00	–0.00	–0.10	–0.11*	–0.18*
Teachers' Qualifications		–0.00	–0.02	–0.04	–0.06	–0.09	–0.12
Teachers' Influence		+0.01	+0.03	+0.04	+0.17*	+0.08	+0.21*
School Climate		+0.02	–0.01	+0.05	+0.02	+0.16*	+0.15*
r^2		0.49	0.36	0.54	0.50	0.48	0.48
	School Climate						
School Size		–0.09*		–0.15*		–0.23*	
Class Size		–0.06		–0.11		–0.06	
Normative Cohesion		+0.25*		+0.29*		+0.31*	
Teachers' Influence		+0.04		+0.05		+0.08	
r^2		0.51		0.58		0.57	
	Teachers' Self-Perceptions of Influence						
School Size		–0.15*		–0.15*		–0.21*	
Normative Cohesion		+0.36*		+0.26*		+0.22*	
r^2		0.18		0.15		0.20	
	Normative Cohesion						
School Size		+0.02		–0.01		–0.10	
r^2		0.07		0.02		0.03	

Table 2. OLS Standardized Regression Weights for School-level Factors Associated with Mean Reading and Mathematics Scores in Public Elementary, Middle, and Secondary Schools (continued)

Factor 1	Factor 2	Elementary (n = 1123)	Middle (n = 496)	Secondary (n = 595)
	Teachers' Qualifications			
School Size		+0.12	+0.11	+0.27
r^2		0.03	0.03	0.09
	Class Size			
School Size		+0.35*	+0.51*	+0.55*
r^2		0.14	0.14	0.38

NOTES: (*) p<.05 based on repeated half-sample standard deviations.

r^2 values include effects of three background factors in addition to factors shown.

accounted for by other factors. Estimates of coefficients for the SEM corresponding to the relations in figure 2 and the specified structural equations are shown in table 3.[26]

If all of the school-based predictors in the OLS regression equations were uncorrelated with each other after background factors were taken into account, then the standardized regression weights displayed in table 2 would, to a first approximation, be the same as the partial correlations displayed in table 1. Furthermore, *if* all of the indicators of each factor in the SEM analysis were perfectly intercorrelated, then the results in table 3 should be approximately the same as those in table 2. (The approximation would not be exact because the SEM equations are estimated jointly, while the OLS equations are estimated separately.) Thus, substantial differences in corresponding coefficients between these tables demonstrate the impact of measurement error and multivariate interactions.

Although SEM is the preferred method of analysis for such analyses because it takes into account the measurement error in predictors, estimation of SEM coefficients requires a large number of degrees of freedom, and the

[26] The corresponding structural equation coefficients for the background factors are not shown in table 3 because they are not germane to the exploration of school-based correlates of achievement. As expected, SEM analyses confirm that all measurement variables are indeed significantly related with the corresponding factors, as indicated in figure 3. See appendix A for results of SEM on the measurement level.

Table 3. SEM Associations of Mean Reading and Mathematics Scores with School-level Factors in Public Elementary, Middle, and Secondary Schools

Factor 1	Factor 2	Elementary (n = 1123)		Middle (n = 496)		Secondary (n = 595)	
		Reading	*Math*	*Reading*	*Math*	*Reading*	*Math*
	Achievement in Reading & Student Mathematics						
School Size		+0.10	+0.02	+0.07	–0.13	+0.50*	+0.18
Class Size		–0.34*	–0.18	–0.37*	–0.11	–0.53*	–0.30
Normative Cohesion		–0.03	–0.00	+0.02	–0.14	+0.05	–0.06
Teachers' Qualifications		–0.02	–0.05	–0.14	–0.17		
Teachers' Influence		+0.03	+0.08	+0.00	+0.18*	–0.09	+0.16
School Climate		–0.30*	–0.30*	–0.06	–0.04	–0.09	–0.14
r^2		0.91	0.66	0.92	0.89	0.89	0.89
	School Climate						
School Size		–0.09		–0.08		–0.33*	
Class Size		–0.12		–0.23		–0.001	
Normative Cohesion		+0.24*		+0.37*		+0.34*	
Teachers' Influence		+0.00		+0.11		+0.00	
r^2		0.66		0.698		0.71	
	Teachers' Self-Perceptions of Influence						
School Size		–0.11		–0.043		–0.21	
Normative Cohesion		+0.39*		+0.35*		+0.30	
r^2		0.21		0.28		0.47	
	Normative Cohesion						
School Size		–0.02		–0.13		–0.19	
r^2		0.10		0.04		0.09	

Table 3. SEM Associations of Mean Reading and Mathematics Scores with School-level Factors in Public Elementary, Middle, and Secondary Schools (continued)

Factor 1	Factor 2	Elementary (n = 1123)	Middle (n = 496)	Secondary (n = 595)
	Teachers'			
	Qualifications			
School Size		+0.16	+0.04	
r^2		0.09	0.11	
School Size	Class Size	+0.54*	+0.75*	+0.70*
r^2		0.37	0.50	0.57
		Statistical Summary Measures		
	GFI (AGFI)	0.95 (0.92)	0.92 (0.86)	0.94 (0.88)
	P2 (d.f.)	547 (122)	473 (122)	400 (93)

NOTES: Entries in table are standardized gamma coefficients.

(*) $p<.05$ based on repeated half-sample standard deviations.

r^2 values include effects of three background factors in addition to factors shown.

resulting estimates can be sensitive to intercorrelations between the indicators of the various factors. Thus, to assess the robustness of the SEM results, it is useful to compare the results from the three approaches. Where they are in conflict, further study of the fit of the data to the assumptions of the method is warranted.

SEM results must be interpreted carefully, however, because the apparent implied causal direction in the regressions is not necessarily valid. Simple interpretation of the results in table 3 requires that one assume that the arrows in the causal model shown in figure 2 imply causality.[27] Any significantly positive or negative coefficient might also reflect a causal relation in the opposite direction.[28] [29]

[27] The selection of factors for exclusion from particular structural equations was based on trial estimation runs in which the goodness of fit index was improved by their deletion.

[28] Table 3 does not include the coefficients for the measurement equations portrayed in figure 3. These coefficients are presented in appendix A in this paper.

[29] Parameter estimates for a variant of the model in figure 2, in which the positions of school climate and achievement are reversed, are given in appendix B in this paper.

A note is necessary concerning the apparently very high percentages of variance accounted for in the reading assessment scores by the SEM analysis (89 to 91 percent). This figure does not indicate the percentage of variance accounted for in the *observed* assessment scores. Because only a single measure of reading assessment was available for each school, the selection of a percentage for measurement error was somewhat arbitrary. However, estimation of the SEM equations did not yield a feasible solution unless measurement error was included and allowed to be correlated between the mathematics and reading scores. Only by setting the measurement error variance to a narrow range of values could a feasible solution be obtained using SAS PROC CALIS. For example, for elementary schools, the measurement error variances for both math and reading scores were set to 0.23. Thus, the school-level factors were accounting for 89 to 91 percent of a factor that itself constitutes 77 percent of the observed school mean reading score variance.[30] Finally, at the high school level, it was necessary to omit the teacher qualifications equation to obtain a convergent SEM solution.

Simultaneously, interpretation of the results presented in tables 1, 2, and 3 provides information about the school-based correlates of achievement and about the extent to which measurement error and correlated predictors can distort analytical results. Through these analyses, we address the following question: "Are school climate, class size, school size, teachers' perceptions of normative cohesion and sense of control, and teachers' experience and education level statistically significantly related to reading and mathematics achievement, based on the SASS student achievement subfile?"[31]

School Climate

The partial correlations in table 1 indicate that a positive school behavioral climate is a correlate of higher average achievement, but only at the middle (for reading) and secondary levels. The OLS regression results in table 2 confirm this finding, but at none of the levels do the SEM results in table 3 indicate a positive relation.

---

[30] A separate series of analyses not reported here explicitly defined achievement as a single factor contributing to two observed measures, reading and math school means.

[31] For the purposes of inferring statistical significance, standard errors of the tabulated estimates were estimated by replicating the analyses on 100 random half-samples of states in the database.

So is school climate a correlate of achievement in middle and secondary schools? Should educators expect that improving school climate might contribute to improved test scores? Examination of the correlates of school climate in tables 1, 2, and 3 sheds some light on this issue. The results from all three sets of analyses suggest that school behavior climates are better in schools with high normative cohesion (i.e., where teachers feel that they have common goals and cooperate) and in smaller schools, especially among students in higher grade levels. Although cohesion and climate are correlated with each other, the partial correlation of achievement with cohesion is smaller than its correlation with climate, so in the secondary school OLS regressions, cohesion becomes a moderator, with a negative coefficient, magnifying the positive coefficient for climate. Two factors in the SEM model eliminate the effect, suggesting that it is an artifact of OLS regression. First, the contribution of normative cohesion to school climate is explicitly taken into account; and, second, the "method" factor representing teachers' positive or negative response tendencies is included.

The message from these analyses, based on the assumptions embodied in figure 2, is that steps to increase normative cohesion among staff may go hand-in-hand with improving school climate, but that these factors are not strong correlates of assessment scores when measurement error is taken into account.

Class Size

As predicted by most current research, reduced *class size* is related to higher academic performance, but the significant relations are primarily limited to reading in this study; the relations with math scores are much weaker. For reading, the partial correlations with class size in table 1 are significant at the middle and secondary school levels; the regression coefficients in table 2 at the middle and secondary level are significant; and the SEM coefficients in table 3 are significant at all three grade levels.

It seems clear from these data that the class size factor is a correlate of reading achievement among middle and secondary schools. At the elementary level, on the other hand, the partial correlation coefficients and OLS results do not corroborate the SEM results. This appears to be an instance where the flexibility of the SEM method enables it to uncover a relation that is hidden by the imperfect relations between observed indicators and underlying factors.

Even in the SEM results, however, class size is a stronger correlate of reading achievement in higher grades. Student/teacher ratios, reported class sizes, and teachers' satisfaction with class sizes all tend to be more favorable in high schools with higher reading (or verbal or language) scores than in high schools with lower average scores.

School Size

The class size and school size factors are positively correlated with each other at all three grade levels, as shown in tables 1, 2, and 3; but the patterns of their correlations with achievement are different. At the secondary level, reading scores are higher in schools with larger enrollments, but at the elementary level there is no significant correlation.[32] The negative partial correlation for middle schools, shown in table 1, is shown to be artifactual in tables 2 and 3: higher reading scores are found in smaller middle schools because those schools have smaller class sizes. Across all three grade levels, there is a tendency (shown in table 3) for school size to be more positively (or less negatively) associated with reading scores than with mathematics scores.

The relation between school size and climate is clearer: at all three levels, the partial correlations and OLS regression results indicate better climates in smaller schools; the SEM results are in the same direction, although they are only statistically significant at the secondary level. Whether smaller schools happen to be in neighborhoods and communities where teachers perceive better school climates or whether smaller schools foster better climates is an issue for further study.

Teachers' Perceptions of Normative Cohesion and Sense of Control and Influence

Normative cohesion and the sense of having influence on school policies and control over classroom decisions are positively related to each other at all three grade levels. While the expectation is that the cultural cohesion establishes a stable foundation for performance, the results in tables 1, 2, and 3 do not indicate that either of these factors is a significantly positive correlate of achievement. In fact, there is a negative relation between normative cohesion

[32] The school size measure is the logarithm of total enrollment.

and achievement in OLS regression at the secondary level, but this is an artifact of the close relation between the normative cohesion and school climate composites. When measurement error in these constructs is taken into account, by SEM analyses, the negative relation disappears, as does the positive relation of school climate to achievement.

Middle and secondary schools in which teachers perceive that they have more than average control over classroom practices and influence on school policies tend to be schools in which mathematics scores are higher. Further study is needed to determine whether this phenomenon represents a specific effect on mathematics teachers or a general "reform" factor (i.e., some school administrators, more than others, recognize both the critical need for math skills and the importance of empowering teachers). The relations among normative cohesion, sense of control, and school climate are consistent with Ingersoll's (1996) conclusion that teachers' autonomy and influence "make an important difference for the amount of cooperation or conflict in schools" (p. 171). Relying on SASS 1987–88 data with very similar measures, he adds that this relationship varies by the locus of teachers' control: when locus is fundamentally social (i.e., "selection, maintenance, and transmission of behaviors and norms"; p. 171), rather than concerned with curriculum and instruction (i.e., "selection of textbooks, topics, materials, and teaching techniques"; p. 171), then the association between teachers' lack of power and conflict in school is strongest.

Overall, there is evidence in the SASS student achievement subfile that organizational characteristics of schools are correlates of student achievement. SASS offers an abundance of opportunity to assess organizational characteristics in American schools (see Baker 1996). Hence, while this report focuses on four organizational characteristics, future analyses of these data might concentrate on such organizational features as organizational inertia (as, for example, in regard to personnel tenure); organizational change (as, for example, in regard to reform issues); or autonomy (as both an intra-organizational feature—room for initiatives—and an inter-organizational feature—dominance of the school district over major decisionmaking issues).

Teacher Qualifications

The final potential correlate of achievement at the school level examined in this study is the average of teachers' qualifications, as represented by years

of teaching experience and attainment of a master's degree. The SASS student achievement subfile did not strongly support the investigation of this factor at the school level, due both to the high level of within-school variance among the responding teachers (i.e., low reliabilities) and to the low intercorrelation between the two indicators (see appendix A in this paper). At the secondary level, the SEM estimation procedure would not even easily converge when these indicators were included. Therefore, the findings of insignificant relations between this teacher qualification composite and reading and mathematics achievement are not unexpected.

Unlike the factors based on teachers' perceptions (school climate, normative cohesion, and sense of control), which can have a communality based on common perceptions of the school as a workplace environment, teachers' experience and education are naturally highly variable within a school, as new teachers are continually added to replace highly experienced teachers who retire. Investigation of these indicators as correlates of achievement appears to require achievement data at the individual classroom level.

It may be that other teacher qualification factors available in SASS, such as matches of college major to teaching assignments and selectivity of the teachers' undergraduate colleges, have sufficient communality within schools to support school-level analyses. Otherwise, study designs such as NELS:88 and NAEP (when specific teacher questionnaires were matched to specific students' performance) are more appropriate for assessing the correlation between teacher qualifications and achievement.

Conclusions

The objective of this report has been to demonstrate and evaluate the strategy of combining a large-scale national survey of schools (SASS), which lacks measures of student achievement, with school-level assessment data from a large number of individual states. If application of this strategy yields new insights about schools or identifies questions that lead to new avenues of research, then its value is demonstrated. If the substantive findings are empty, the strategy is less attractive.

To demonstrate the strategy, a set of 18 composites of SASS data, including student background information, organizational information, teachers' qualifications, and school climate perceptions, were constructed and merged

with school reading and mathematics mean scores. The resulting data were analyzed using correlational, multiple regression, and structural equation model analyses. These analyses only begin to tap the richness of the SASS database: selection of other subsets of the SASS data or other analytical methods could add to the evaluation of the strategy.

Substantive Findings

The clearest result with respect to correlates of achievement is that reading scores are higher in schools with smaller class sizes. This result, obtained from structural equation modeling using both state assessment data and NAEP adjustments for between-state variance in achievement, is consistent across grade levels (see table 3). While there are alternative causal explanations for this finding, such a finding in a large sample of public schools in 20 states is an important corroboration of the controlled research results that indicate that class size makes a difference.

The positive relation between small classes and reading scores was stronger for secondary schools than for elementary schools. In secondary schools, the positive association with reading included both large schools and small classes. The relation between class size and achievement was specific to reading scores; it was much weaker for mathematics.

Substantive findings were not limited to class size. Teachers' perceptions that they had control of classroom practices and influence on school policies tended to be greater in middle schools, and possibly in high schools, in which mathematics scores were higher. The analyses carried out do not provide a causal explanation for this relation, but its statistical significance suggests a potentially fruitful area for both additional studies of SASS measures and controlled research.

Because the data are not longitudinal, causal inferences must be treated much more tentatively than conclusions based on data on the achievement gains of a specified set of students over time. Also, because the data are school means, they cannot address the factors that differentially affect the achievement of different students in the same school. Nevertheless, findings from analyses of the SASS student achievement subfile, based on over 2,000 schools in 20 states, can contribute to the overall educational policy database.

Methodological Findings

The primary conclusion reached in this study is that because the data are readily available, the strategy of matching school-level assessment scores to a national survey is feasible, and not costly, and leads to valid and reliable conclusions about correlates of public school achievement across much of the United States. The additional step of linking the database to State NAEP to capture between-state achievement variation provides additional informational value and is also feasible and not costly. In a separate report (McLaughlin and Drori 1999), a comparison between analyses that include the State NAEP adjustment and simple pooled within-state aggregations indicated that some correlates were stronger (e.g., between class size and achievement) when the State NAEP adjustment was used.

A positive methodological finding was the generalizability of the between-state achievement measures across grade levels. Although state assessment scores were available for grades from 3 to 11, NAEP reading scores for individual states were only available for grade 4. If the ordering of states in reading achievement changed substantially from grade 4 to grades 8 and 11, then the results of overall analyses of middle school and high school data would be diluted by linkage error.

The extension of the NAEP adjustment proved valid, in that the findings for secondary schools are as meaningful as the findings for elementary schools. This conclusion is not surprising, given the very high correlation of State NAEP means in different grades and subjects, but its support in this study may suggest new uses of State NAEP data in conjunction with state assessment data.

A limitation on the validity of aggregating teacher data for school level analyses became apparent in the findings concerning teacher qualifications (average years of teaching experience and percent having a master's degree). These measures, unlike the teachers' responses to questions about school policies and school behavioral climate, had very low reliability *as measures of the school,* because there was relatively little systematic between-school variation: most of the variation was between teachers at the same school. This problem was manifest in the low intercorrelation between these measures; as a result, the analytical findings concerning the relations of this teacher qualifications factor were uninterpretable. In fact, the SEM software had difficulty even

converging to a solution when this teachers' qualification factor was included in structural equations.

Finally, the procedure of examining ordinary least squares regressions and partial correlation coefficients to understand the results of structural equation model analyses proved valuable, in that it suggested explanations for unusual findings, such as the negative coefficient for normative cohesion in predicting middle school achievement. Although the data are purely correlational, there are logical constraints, such as that school factors probably do not cause differences in student background characteristics in the short term ("white flight" notwithstanding). Interpretation of the results of structural equation modeling in terms of hypothetical path models, such as those shown in figure 2, can lead to fruitful suggestions for avenues of research and policy development.

Future Research

Two broad areas of research stemming from this study appear to be fruitful: development of a measure of a school's achievement gains over time which can be associated with SASS measures and further refinement of the linkage functions between state assessments and NAEP.

Every school addresses the needs of a different student population with different resources, and it is therefore unfair to hold all schools accountable to the same achievement standard. However, a number of states are turning to reform criteria that base decision-making on measures of *gains* in achievement over years. Although SASS cannot easily add longitudinal student growth data, it is certainly feasible to add other years' school-level achievement data to the subfile. Specifically, the addition of 1997–98 reading scores, linked to the 1998 grades four and eight State NAEP reading assessment and CCD data on changing enrollment patterns and resources over the intervening years, would provide the basis for identifying SASS factors (measured in 1994) that are predictive of gains in achievement. For example, one wonders whether staff turnover rates would portend gains, other things equal. Of course, states continue to develop and refine assessment systems, and the state assessment test scores for a school in 1998 may not be equivalent to scores obtained in 1994, so linkage of measures of achievement gains over time to repetitions of State NAEP is an essential requirement for the development of a longitudinal database.

The power of the database for longitudinal analyses can be greatly enhanced with the addition of the next cohort of SASS. If a subsample of schools included in SASS in 1994 are also included in 1999–2000, then using the 2000 State NAEP assessment for adjustment of mathematics scores would enable the matching of longitudinal changes in SASS school-based factors with longitudinal changes in achievement, controlling for longitudinal changes in student background factors.

A second line of research would focus on improving the achievement measures included in the SASS student achievement subfile. The linkages used for the analyses presented in this report were based entirely on the means, standard deviations, and correlations between State NAEP and state assessment school means. The errors in these linkages can be diminished significantly by more detailed analysis of the relations among the scores. In particular, current research for the National Center for Education Statistics has found that linkages to NAEP can be improved by considering nonlinear terms and by including demographic indicators. For example, all state reading assessments are sensitive to racial-ethnic differences, but some are more sensitive than others. Their sensitivities could be matched to NAEP's measurement of the distribution of racial-ethnic achievement differences by explicitly including that matching factor in the NAEP adjustment step in constructing the SASS school-level achievement score. The result would be increased comparability of within-state variation in the achievement measure across states.

References

Baker, D. P. (1996). Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with Comments on School Reform, Governance, and Finance. In J. Mullen and D. Kasprzyk (Eds.), *The Schools and Staffing Survey: Recommendations for the Future* (NCES 97–596) (pp.19–37). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Bollen, K. A., and Bollen, W. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Boruch, R. F., and Terhanian, G. (1996). So What? The Implications of New Analytic Methods for Designing NCES Surveys. In G. Hoachlander, J. Griffiths, and J. Ralph (Eds.), *From Data to Information—New Directions for the National Center for Education Statistics* (NCES 96–901). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: Government Printing Office.

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

Herriot, R. F., and Firestone, W. A. (1984). Two Images of Schools as Organizations: A Refinement and an Elaboration. *Educational Administration Quarterly 20*(4):41–58.

Ingersoll, R. M. (1996). Teachers' Decision-making Power and School Conflict. S*ociology of Education 69*(2):159–176.

Kaufman, P. (1996). Linking Student Data to SASS: Why, When, How. In J. Mullens and D. Kasprzyk (Eds.), *The Schools and Staffing Survey: Recommendations for the Future* (pp. 53–65). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

McLaughlin, D. H. (forthcoming). *Study of the Linkages of 1996 NAEP and State Mathematics Assessments in Four States* (NCES 2000–461). National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

McLaughlin, D. H., and Drori, G. (forthcoming). *School-level Correlates of Academic Achievement: Student Assessment Scores in SASS Public Schools* (NCES 2000–303). National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Mullis, I. V. S., Campbell, J. R., and Farstrup, A. E. (1993). *NAEP 1992 Reading Report Card for the Nation and the States.* (No. 23-ST06). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1993). *NAEP 1992 Mathematics Report Card for the Nation and the States* (No. 23-ST02). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1996, June). *Profile of 1994–95 State Assessment Systems and Reported Results*. Washington, DC: Author.

Reese, C. M., Miller, K. E., Mazzeo, J., and Dossey, J. A. (1997*). NAEP 1996 Mathematics Report Card for the Nation and the States* (NCES 97–488). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Roeber, E., Bond, L., and Braskamp, D. (1997*). Annual Survey of State Student Assessment Programs: Fall 1996.* Washington, DC: Council of Chief State School Officers.

Roeber, E., Bond, L., and Connealy, S. (1998*). Annual Survey of State Student Assessment Programs: Fall 1997.* Washington, DC: Council of Chief State School Officers.

Wu, G., Royal, M., and McLaughlin, D. (1997). *Working Paper: Development of a SASS School-level Student Achievement Subfile, Using Existing State Assessment and State NAEP Information* (NCES 97–44). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Appendix A
Results of SEM: Measurement Level

A measurement model is required to link the latent variables in the structural equation model to SASS measures. Four tables provide the information necessary for understanding the latent variables.

Table A displays the indicators of each latent trait and the SASS variables used in its computation. For example, poverty is measured by the ratio of the reported number of free-lunch eligible students (S1655 and S1660) to the total enrollment (S0255), and by the average of teachers' responses to the item asking whether poverty is a slight, moderate, or serious problem at the school (T1165).

Table B gives a reliability estimate for the teacher-based indicators, that is, a measure of the tendency of teachers at the same school to give the same responses. The estimate is one minus the ratio of (a) the sum of within-school variances, divided by the total number of teachers responding, to (b) the variance of school means. Values substantially less than 0.5 (such as average class size, classroom control perceptions, and years of experience) indicate that more of the variance in the indicator is within schools than between schools, and alternative indicators should be given greater weight in the model.

Table C gives intercorrelations between the indicators. For example, the lowest intercorrelations (0.28 and 0.29) are for the average class size with the other two indicators of the class size factor and between average years of teachers' experience and percent of teachers with a master's degree, all at the elementary school level. These intercorrelations need not be substantial for SEM analyses, because the estimation procedure should identify the weighting of the indicators that most effectively accounts for variance in other measures. Nevertheless, values substantially less than 0.50 indicate that most of the variance in the indicators is not in common across the latent trait.

Table D contains the SEM measurement parameter estimates which were obtained simultaneously with the structural equation parameter estimates. For example, the latent poverty variable is set to be in the same units as the free-lunch eligible fraction by presetting its coefficient to 1.0; at the elementary school level, 0.65^2 or 42 percent of the variance in the free-lunch-eligible fraction is attributed to the latent poverty trait.

Table A. School-level Achievement Correlates: Factors, Indicators, and SASS Items Used

Category	Factor	What It Means	How We Measured It Our Indicators	SASS Items Used
Student Background				
	Poverty	Low SES	Ratio of lunch program eligibility, and teachers' identification of a poverty problem	S1655 S1660 (by S0255) T1165
	Race	Racial tension and problems	Ratio of white students, and teachers' identification of a problem of racial tension	S0425 (by S0255) T1170
	English Language proficiency	Low proficiency of English	Ratios identified as having language problems (i.e., ratios of participation in English enhancing programs), ratio of Hispanic students, and teachers' identification of limited English as a problem for students	S1410 S1295 S0415 (by S0255) T1190
Teaching Quality				
	Teacher qualifica-tions	Teachers' degrees and teaching experience	Professional education (i.e., having a master's degree) and years of teaching experience	T0235 T0095 T0100 T0105 T0110
School Climate				
	School atmosphere and problems	Student behavior	Teachers' identification of problems with (set #1) tardiness, dropping out, lack of academic challenge, vandalism, drug abuse, physical conflicts, verbal abuse of teachers, physical attacks on teachers, and teacher absenteeism, and(set #2) student absenteeism, cutting class, apathy, robbery or theft, disrespect of teachers, alcohol abuse, and weapons in school	T1005 T1075 T1080 T1085 T1100 T1105 T1115(5)T1150 T1325 T1330 T1340 T1345

Table A. School-level Achievement Correlates: Factors, Indicators, and SASS Items Used (continued)

Category	Factor	What It Means	How We Measured It Our Indicators	SASS Items Used
Organizational Features				
	School size	Organization size	Number of students enrolled	S0255
	Class size	Resource per student; crowding	Average class size, ratio of enrollment to number of teachers, and teachers' sense of satisfaction with class sizes	S0255 T0830-T0970 S0255 S0910 S0850; T1285
	Teacher Influence	Teachers' control over school policies and classroom arrangements	Teachers' sense of influence over such school matters as setting discipline policy, determining content of inservice programs, hiring, school budget, teacher evaluation, and establishing curriculum; and, teachers' sense of influence over such classroom matters as selecting textbooks and other instructional materials, selecting content, topic, and skills to be taught, selecting teaching technique, evaluating and grading, disciplining students, and determining amount of homework assigned	T1015 T1020 T1025 T1030 T1035 T1040 T1045 T1050 T1055 T1060 T1065 T1070
	Normative cohesion	Clarity of norms	Colleagues share beliefs and values; principal enforces rules and backs up staff; receive support from parents; principal lets staff know expectations; goals and clear; principal knows what he/she wants and communicates to staff; behavior rules are consistently enforced.	T1200 T1225 T1245 T1255 priorities are T1260 T1265 T1295
		Cooperation among school staff	Coop among staff; whether teachers coordinate course with other teachers; sense that administration-staff relations are supportive and encouraging; teachers integrate library/media sources into curriculum	T1205 T1270 T1290 T1310

NOTE: SASS items beginning with "T" are on the teacher questionnaire, and those with "S" are on the school questionnaire.

Table B. Reliability Coefficients for Teacher-based Components of School-level Factors, by School Level (Estimated fraction of sample mean variance that is between schools)

		Elementary	Middle	Secondary	All Schools
Class Size	Average	.31	.37	.44	.42
	Satisfied?	.45	.47	.54	.51
Climate	Climate 1	.70	.74	.77	.82
	Climate 2	.78	.78	.86	.85
Normative Cohesion	Cooperation	.51	.46	.47	.50
	Clear Norms	.54	.55	.61	.60
Teacher Control	Classroom Control	.36	.38	.50	.41
	Influence on School Policies	.61	.62	.62	.63
Teacher Qualifications	Years Experience	.26	.14	.27	.24
	Masters	.41	.39	.44	.41
Poverty	Problem?	.82	.79	.82	.80
Minority Conflicts	Problem?	.73	.79	.84	.80

Table C. Intercorrelations of Components of School-level Factors, by Level

		Elementary	Middle	Secondary	All Schools
Class Size Factor					
Average Class Size	Student/ Teacher Ratio	.28	.50	.57	.45
Average Class Size	Class Size Satisfaction	.29	.45	.49	.40
Student/ Teacher Ratio	Class Size Satisfaction	.43	.42	.55	.48
Climate Factor					
Climate 1	Climate 2	.61	.72	.50	.72
Normative Cohesiveness					
Cooperation	Clear Norms	.66	.67	.66	.69
Teacher Control and Influence					
Classroom Control	School Policy Influence	.50	.49	.49	.48
Teacher Qualifications					
Years Experience	Masters	.29	.30	.37	.32
Poverty					
Pct. Free-Lunch Eligible	Poverty a Problem	.55	.51	.47	.49
Racial/Ethnic Minorities					
Pct. Minority	Race Conflicts a Problem	.49	.45	.43	.45
Language Minorities					
Pct. Limited English	Pct in ESL Classes	.80	.75	.70	.77

Table D. School-level Factors: SEM Measurement Model Estimates

Indicator	Elementary	Middle	Secondary
Pct. free lunch eligible	=1.0 (.65) Poverty+ .76e	=1.0 (.66) Poverty+.75e	=1.0 (.61) Poverty+.79e
Teachers see poverty a problem	=1.26 (.81)* Poverty -.34 (-.22)* T_Quest+ .54e	=1.13 (.70)* Poverty -.70 (-.43)* T_Quest+ .57e	=1.38 (.86)* Poverty - .11 (-.07) T_Quest+ .50e
Pct. Limited English Proficient	=1.0 (.87) Language+ .49e	=1.0 (.89) Language+ .46e	=1.0 (.85) Language+ .53e
Pct. in ESL instruction	=1.05 (.90)* Language+.44e	=.94 (.81)* Language+.58e	=.98 (.82)* Language+ .57e
Pct. minority	=1.0 (.81) Race+ .58e	=1.0 (.85) Race+ .52e	=1.0 (.83) Race+ .56e
Teachers see racial tension a problem	=.76 (.63)* Race -.48 (-.30) T_Quest+ .71e	=.62 (.50)* Race -1.01 (-.62)* T_Quest+ .61e	=.64 (.54)* Race -.17 (-.11)* T_Quest+ .84e
Years teaching experience	=1.0 (.57) T.Qualif +.82e	=1.0 (.80) T.Qualif+ .60e	=1.0 (.51) T.Qualif+ .86e
Masters degree	=.90 (.51)* T.Qualif+ .86e	=.46 (.37)* T.Qualif+ .93e	=1.42 (.73)* T.Qualif+ .68e
Perceived classroom control	=1.0 (.58) T.Control+ .82e	=1.0 (.59) T.Control+ .80e	=1.0 (.74) T.Control+ .68e
Perceived policy influence	=1.51 (.87)* T.Control+.49e	=1.39 (.83)* T.Control+ .56e	=.90 (.67)* T.Control+ .75e
School size	=1.0 (1.0) School.Size+ .0e	=1.0 (1.0) School.Size+ .0e	=1.0 (1.0) School.Size+ .0e
Average class size	=1.0 (.39) Class.size+ .92e	=1.0 (.61) Class.size+ .80e	=1.0 (.72) Class.size+ .70e

Table D. School-level Factors: SEM Measurement Model Estimates (continued)

Indicator	Elementary	Middle	Secondary
Student/teacher ratio	=1.78 (.70)* Class.size+.71e	=1.31 (.80)* Class.size+ .60e	=1.12 (.81)* Class.size+ .59e
Teachers see class size as satisfactory	=1.53 (.58)* Class.Size + 1.0 (.61) T_Quest+ .54e	=.87 (.45)* Class.Size + 1.0 (.54) T_Quest+ .71e	=.86 (.63)* Class.Size + 1.0 (.64) T_Quest+ .43e
Clarity of norms	=1.0 (1.0) Norm.Co+ .0e	=1.0 (1.0) Norm.Co+ .0e	=1.0 (1.0) Norm.Co+ .0e
Cooperation	=.65 (.65)* Norm.Co+ .76e	=.66 (.65)* Norm.Co+ .76e	=.65 (.65)* Norm.Co+ .76e
School behavioral climate 1	=1.0 (.94) School.Climate -.33 (.21)* T_Quest+ .28e	=1.0 (.85) School.Climate -.78 (.49)* T_Quest+ .17e	=1.0 (.97) School.Climate -.18 (.11)* T_Quest+ .21e
School behavioral climate 2	=.98 (.92)* School.Climate -.35 (-.23) * T_Quest+ .31e	=.93 (.79)* School.Climate -.79 (-.50) * T_Quest+ .34e	=.91(.91)* School.Climate -.27 (-.18)* T_Quest+ .38e

NOTES: Data in table include: Unstandardized coefficient or factor loading (standardized coefficient or factor loading) significance level, factor name, and magnitude of error term. Indicators with an unstandardized factor loading of 1.0 are the reference indicators. Error terms of .0 were set to this value in order to identify the model.

Appendix B
Factors Associated with School Climate and Achievement in Public Schools, Reversing the Causal Order between These Two Factors: Results from SEM Analyses

A separate SEM analysis was carried out in which the school climate trait was omitted from the equation for achievement and the school achievement trait was included in the equation for school climate. (This analysis was carried out using a single school achievement trait, measured by the two indicators of average reading and mathematics scores.)

Independent Factor	Dependent Factor	Elementary (n = 1124)	Middle (n = 496)	Secondary (n = 595)
	Student Achievement			
School Size		+0.10	+0.01	+0.39*
Class Size		–0.25*	–0.27	–0.37*
Normative Cohesion		–0.09*	–0.02	–0.00
Teachers' Qualifications		–0.01	–0.13	–0.12
Teachers' Influence		+0.02	+0.02	–0.05
r^2		0.69	0.85	0.81
School Size				
	School Climate	–0.05	–0.06	–0.22*
Class Size		–0.20	–0.26	–0.09
Normative Cohesion		+0.22*	+0.37*	+0.34*
Teachers' Influence		+0.01	+0.11	+0.02
Student Achievement		–0.29	–0.03	–0.11
r^2		0.68	0.75	0.70
	Statistical Summary Measures			
	GFI (AGFI)	0.95 (0.92)	0.91 (0.85)	0.93 (0.88)
	P2 (d.f.)	606 (130)	552 (130)	506 (130)

NOTES: Entries in table are standardized gamma coefficients.

(*) p<.05 based on repeated half-sample standard deviations. The same factors and equations were included as represented in table 6, except for the reversal of achievement and school climate.

r^2 values include effects of three background factors in addition to factors shown.

Response: Opportunities for Design Changes

Valerie E. Lee
University of Michigan

I have had much experience in the statistical analysis of NCES data sets, particularly those with a longitudinal design. I feel very grateful to NCES for collecting these excellent data and making them available to researchers. Imagine the cost if we had to collect such data ourselves! However, in conducting many studies, it is not unusual for researchers to discover many difficulties—and to want to change future data collections to avoid such difficulties.

Thus, I am using this conference, and my participation here, as a public opportunity to encourage NCES to consider some changes in their data collection designs. I believe that the implementation of these suggested changes would help researchers address policy-relevant questions about education more accurately and reliably. I use my comments about the papers in this session as a "launching pad" for making these suggestions.[1]

Comments on the Brewer and Goldhaber Paper

The Brewer and Goldhaber paper, "Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88," is a solid, even classic, example of how economists conduct educational production function research. The dependent variable these researchers explore is a measure of learning, which they conceptualize as the change in students' achievement over a two-year period—the first two years of high school. The authors focus on achievement gains in the four subjects that were tested in NELS. I agree with the approach that uses gains in achievement in individual subjects as measures of learning.

---

[1] The opinions expressed in this paper are those of the author. Address any comments to Professor Valerie Lee, School of Education, University of Michigan, 610 East University Avenue, Room 4220, Ann Arbor, MI 48109.

The Research Focus

Brewer and Goldhaber's research focuses on estimating achievement gains as functions of classroom conditions and qualifications of teachers. Their analyses take typical control variables into account. The researchers capitalize on the NELS data collection strategy, in which data collected from two of each student's teachers may be matched directly to each student in the subject in which the teacher has taught the student in his or her tenth grade year. In their analyses, these researchers use a single unit of analysis—the student—and they use ordinary least squares (OLS) regression as their major analytic method. From my vantage point, the research described by Brewer and Goldhaber is characterized by a few methodological difficulties, some of which the authors could not have avoided. As stated, I use these difficulties to highlight some shortcomings of NELS data collection designs. I recognize that economists, such as these authors, may have several differences of opinions with sociologists of education like me. Nonetheless, let me spell out two major difficulties.

Design Difficulty Number One

By intent, the NELS data are not designed so that several students are "nested" within each teacher. Moreover, even if several NELS students do have the same teacher for the same subject at tenth grade (i.e., they are matched to the same teacher in this data set), the students are not necessarily in the same classes even when they have the same teacher. To me, the most logical mechanism through which a teacher's qualifications might be linked to a student's learning would be that the teacher's qualifications would somehow influence how he or she teaches, i.e., through instruction. To oversimply this mechanism, let us assume that instruction is to be broken down into two components. One component is related to the content the teacher imparts. Content would take place at the level of the individual class, and teachers would vary their content depending on the class they were teaching (and the students in the class). The second component is the teacher's expertise. This component is one that the teacher would "carry" with him or her regardless of the class being taught; it might be constant across the students (and classes) taught. In the analyses described by Brewer and Goldhaber, and fundamental to the structure of the NELS data, we must focus only on the second component of qualifications: subject matter expertise. The first component, pedagogical or instructional variation by class, must be ignored.

Therefore, with NELS, we are unable to differentiate between the sort of instruction teachers actually use and the expertise they bring to the subject. With this data structure, where there are very few teachers per student and not all the students taught by a single NELS teacher are actually in the same class, we are unable to either theoretically specify or empirically differentiate whether teachers' qualifications—a major independent construct in this research—actually influence student learning through what material is taught, i.e., content, or how material is taught, i.e., expertise level. To me, education is fundamentally about such things. However, the structure of NELS data limits researchers' ability to tease out the differences between content expertise and pedagogical expertise within the construct of "teachers' qualifications." I do not think there is a single thing Brewer and Goldhaber could do to solve this vexing problem. I suggest, however, that they could have recognized it in their paper.

Solution to Design Difficulty Number One

I believe there is a solution to this problem, and it has to do with changing the basic data collection design. If more students were sampled per school, even at the cost of sampling fewer schools, this important problem would be solved. With such a design, we could recognize that students are in fact "nested" both within classrooms as well as within teachers, and that teachers' basic instruction *does* change across the classes they teach. This is not the first time I have made this suggestion to staff at the National Center for Education Statistics. However, I do not want to lose this golden opportunity to say it once again—that NCES should actually include multiple students in particular classrooms as part of their sampling strategy. In this way, we researchers would be able to investigate the nested nature of schooling at three levels: students, nested in classrooms, which are in turn nested in schools. That is, we would be able to tease out these important levels of variability in the nature of schooling.

Design Difficulty Number Two

The achievement outcome that these researchers use—change over time on the same test—has been used by many of us very frequently. As mentioned, I see this as a measure of learning, specifically growth in achievement in particular subject areas between the eighth and the tenth grades. This way of measuring learning capitalizes on the value of longitudinal research, something that is surely not lost on the participants in this conference. It is hard to think of anyone who would not laud this feature of NCES databases as cru-

cially valuable. Despite the value of longitudinal data, there is a basic design flaw endemic in this type of research. Brewer and Goldhaber link learning between eighth and tenth grade—a two-year period—to the qualifications of the teachers to whom the students were exposed in tenth grade only. The obvious problem remains: What happened in the ninth grade? At one level, one could conclude from this paper that Brewer and Goldhaber's findings probably represent an *underestimate* of relationships that might actually be there, simply because the authors were looking at only part of the students' educational experiences over the two-year time period captured by the achievement gain. Further, high schools students could actually have been in these teachers' classes for only a single semester.

Solution to Design Difficulty Number Two

The obvious solution is to collect achievement and instructional information every year. That is, if we want to link instruction to learning using NCES data, we need to match the data collection period to the instructional period; such a match represents a basic issue of construct validity. If we wish to investigate links between the character of instruction and students' learning as a result of that instruction—a fairly basic question in educational research—then we need an appropriate data structure to be able to do so. The NELS structure, with biennial data collection, does not allow us to be able to do so. Unfortunately, this design means that most of the research conducted with NELS data does not really look at instructional effects. One possibility for researchers is to capitalize on the fact that many school districts in the United States test all students annually. For example, I have conducted a couple of studies using an excellent data set collected in Chicago, where the district tests every student every year. These district data are linked to data collected through periodic surveys conducted by the Consortium for Chicago School Research. The excellent cooperation between the Consortium and the Chicago Public Schools has allowed researchers to link survey data to annual test data. As a result, analysts may measure change in achievement over one year on the same test, and these achievement gains may be linked to survey data from students and teachers about students' educational experiences during the same year. The Consortium didn't have to collect achievement data themselves; it was a major political accomplishment to be able to combine data of this type. I suggest that if this has already been accomplished in Chicago, perhaps it can happen at the national level. Rather than constantly collecting new data, we should at

least think about how to capitalize on existing test data (and equate scores across different tests), as there is surely enough testing going on in the world.

Summary of Comments on the Brewer and Goldhaber Paper

I agree with these authors on three issues and disagree with them on three other issues. First, I agree with the authors that NCES decisions to collect data from teachers that can actually be linked to students are a real advance over previous NCES data collection efforts in NELS. I would like to think of that advance as the first stage in a data collection design that could actually be vastly improved. The first stage is surely important. Second, these authors and I also agree on the need for a better sampling design. In their paper, they argue that having more students sampled per school, and possibly sampling by classroom, would be useful. Third, I also agree with them about the potential usefulness of being able to link data that are collected from SASS with NELS (SASS is a cross-sectional NCES data collection effort.) SASS has information about so many schools and teachers. At present, we cannot link these two datasets, because each has sampled different sets of schools. In the future, it would be useful to have the longitudinal data collections occur in schools that are also part of SASS.

I disagree with Brewer and Goldhaber on three other issues. The first is their use of a single unit of analysis and a methodology restricted to single-unit analyses: OLS regression. That would be acceptable if we knew that a very low proportion of variability in the outcome occurred anywhere but between students. However, other researchers have shown that with the NELS achievement tests, perhaps 25 to 30 percent of the total variability in these test scores lies systematically *between schools*. The authors ignored this. Therefore, they assumed that there was no variability between schools. This is a serious oversight to me. I teach courses in hierarchical linear modeling (HLM), a methodology that is meant to address the multilevel nature of educational data. I feel a strong need to state what may seem obvious: multilevel questions call for multilevel methods.

A third issue about which I disagree with Brewer and Goldhaber is their lack of attention, in their paper and in their brief summary statement in their presentation, to comparisons between public and private schools in other research they have done. I found the simple mention of this type of comparison,

without giving much detail about it, to be quite problematic. I have, with several colleagues, conducted a substantial amount of research on cross-sector comparisons. I would like to have seen what they did, and I would like to have been able to comment on it. Presenting findings in a very summary form, without making them transparent enough so that readers can figure out the analyses from which those summary findings have come, presents a problem. In papers like this one, authors should provide sufficient detail for the benefit of interested readers. I suggest that otherwise such issues don't get raised at all. Some questions that come to mind include: "There are different types of private schools, but did the authors take that into account?" and "What about the social distribution of achievement—a measure of social equity in schools?" These are questions I have examined in some detail in research comparing public and private schools. Other researchers have done this type of research as well.

Comments on the McLaughlin and Drori Paper

The second paper in this session, "School-level Correlates of Reading and Mathematics Achievement in Public Schools," is the work of McLaughlin and Drori. As I see it, the major purpose of the research described in this paper is to demonstrate to the research community that it is possible to merge data from different sources, collected for different designs and for different purposes, into a single data source.

Two Ways to View This Paper

Although it was not exactly clear how McLaughlin and Drori would like readers to view their paper, there are two rather different ways that readers could actually make use of this work. In one approach, readers would regard the analyses heuristically. Viewed through this lens, the authors would like us to see their work as an example of what could be done, i.e., how different NCES data sets could be linked. A major research question taking the heuristic approach might be as follows: "Can researchers really make use of data that are merged in this way?" Using a second lens, readers would consider the paper's substantive analyses and findings. This lens would result in research questions of the following type: "In which types of school do students achieve at higher levels?" Personally, I am more comfortable with the heuristic approach. Given the methodological and conceptual difficulties, I am cautious about the validity of findings and any substantive conclusions drawn from them.

A Problem of External Validity

Both in the presentation and in the paper on which it was based, Don McLaughlin provided much detail about the sample and the methodology used for linkage. Therefore, those details do not need repeating. There are, however, a few issues I would like to mention. One issue is the potentially differential weighting of data by state. For example, the number of sampled schools per state ranged widely: between 50 and 335. Moreover, these numbers are not always proportional to the population. Given that the analyses are at the school level, that kind of between-state variation leads to weighting some states up considerably, while weighting down others. As I see it, this differential weighting arises from two sources: (1) the differential correlations of some state-level information with data from NAEP and (2) the fact that the researchers have a substantial amount of information from some states and very little from other states. Without taking the systematic differences in data quality by state into account, it is risky to generalize these results even to the 20 states for which data are available. Let us consider the outcome variable: school average achievement. The authors first created within-state school averages of students' standardized test scores in several subjects, and they then introduced between-state adjustments using information from NAEP. As McLaughlin mentioned, the authors also weighted these scores by the degree to which the state assessments were correlated with NAEP. The weighting varied considerably—between 0.37 and 0.86. This resulted in schools' average achievement scores in some states (e.g., Kentucky) being weighted down quite substantially. I am not trying to dwell on the specifics here. Rather, I am raising an issue of external validity. For whom, really, are the results to be generalized?

Unit of Analysis, Revisited

In discussing the Brewer and Goldhaber paper, I raised the issue of the appropriate unit of analysis. In that paper, the researcher used a single-level analysis at the level of students to attempt to assess how teacher quality influences students' learning. I suggested that such issues are more appropriately addressed with multilevel methods such as HLM. This same issue—deciding on the appropriate unit (or units) of analysis—is also relevant for the McLaughlin and Drori paper. Whereas Brewer and Goldhaber assumed in their analysis that all of the variance in achievement was between students (although we

know that 25 to 30 percent of the variance in achievement lies systematically between schools), McLaughlin and Drori have, by conducting school-level analyses, assumed that all of the variance lies between schools, i.e., they have ignored the large proportion of variance that lies between students in the same schools. In more technical terms, the first set of authors assumed an intra-class correlation in the dependent variable of 0, and the second set of authors have assumed that the intra-class correlation in school achievement is 1. Neither assumption is correct. Brewer and Goldhaber, as I mentioned, had systematically ignored the 25 to 30 percent of the variance in test scores that was between schools. On the other hand, McLaughlin and Drori have ignored the 70 to 75 percent of the variability that is between students within schools.

A Few Other Issues

A few other issues about the McLaughlin and Drori paper relate to the difficulty in drawing substantive conclusions from their results. One issue relates to aggregation bias. These authors have used only school-level variables, many of which they have created through aggregating student-level data. Quite simply, aggregated variables do not typically have the same meaning as the individual variables from which they are created. For example, school average SES may measure the types of students who attend a school, but it also measures resource availability in the school. Yet it doesn't measure the social homogeneity of the student body (which the individual-level measure would capture). Another relevant issue is model specification. The original model with which the authors introduced the paper was very broad. Moreover, there were multiple variables introduced to operationalize each of the constructs they wished to examine. In my opinion, the original model was too broad. The fact that a large number of variables were introduced into the analysis, but only a few "survived," suggested an approach typified by the following: "Well, let's see what counts?" I prefer causal models that are derived theoretically, that are more focused in scope. I admit that each of the constructs was discussed in a theoretical context in the early part of the paper. Still, the model was too broad.

A final issue I would like to discuss relates to drawing causal conclusions from cross-sectional data. The authors were, in fact, quite careful throughout the paper to introduce multiple disclaimers about not conducting causal analysis. However, by its very nature, the type of analysis conducted here is causal. The authors selected a dependent variable (average school achievement) and a

large number of independent variables. I believe that we are in fact conducting causal analyses whenever we employ typical general linear model methods, regardless of the disclaimers we introduce. The findings, moreover, were not always stable. Some independent variables were statistically associated with school average achievement using one analysis technique, e.g., structural equation modeling, but not with another technique, e.g., OLS regression. Which set of results should we believe?

A Contradictory Conclusion

The authors of these papers surely did not expect that the two papers would be compared. However, comparison seems almost inevitable, since I've been asked to comment on both of them. Again, the issue is as follows: "Which results do we believe?" Both sets of authors investigated whether teachers' qualifications influence student learning. In their paper, Brewer and Goldhaber concluded that teachers' qualifications count. Although McLaughlin didn't mention it in his oral presentation, in his paper he reached the conclusion that teacher qualifications do not count. I realize that the two papers used different data, used different approaches, and examined the question at different units of analysis. Yet both used data that are at least meant to be nationally representative, and both addressed the same overarching research question. As readers interested in drawing substantive policy conclusions from quantitative research, what are we supposed to learn from two papers with divergent conclusions?

Final Words

Let me end by briefly summarizing the issues I wish to emphasize. Both papers have posed what are essentially multilevel questions, but neither has used multilevel methods. Both are essentially posing questions about school effects: "How do characteristics of schools and the classrooms in them influence the learning of students educated in those schools and classrooms?" Such questions require the use of multilevel methods. A second major issue is how we should measure the effects of instruction on learning. I have offered some suggestions here about more appropriate designs for data collection. If we want to link data on teachers and teaching to student achievement and learning, we need data on more students in each school, probably a sampling design where students are systematically nested within teachers and classrooms. This is important. In my opinion, the "action" in education really is right there in the classroom. So let us

get good data so that we can actually explore education at this level. I do not mean to be ungrateful to NCES; they have been helpful to me. Nonetheless, in the spirit of this conference, I think improvements can be made, our methods can be done better, so that is what I am saying. Let us do them better.

Section IV.
Policy Perspectives and Concluding Commentary

Assessment Trends in a Contemporary Policy Context

Marshall S. Smith
Under Secretary of Education
Acting Deputy Secretary of Education
U.S. Department of Education
Washington, DC

For the last 30 years, the "gap" between the scores of African American and white students on standardized tests of reading and mathematics has been a thorny and controversial issue. Efforts to understand—and reduce—the gap have highlighted the challenge of simultaneously making progress toward the two co-equal guiding goals of American public education: educational excellence and educational equality.

In terms of both methodology and research perspectives, the papers presented in this seminar book make thoughtful contributions to our understanding of the achievement gap. These papers, along with those in the Jencks and Phillips (1998) volume, enrich our understanding of policies and programs that will raise overall student achievement while, at the same time, helping to close the achievement gap between black and white students.

Putting such policies in place is a primary focus of the legislative proposal that President Clinton sent to Congress for the reauthorization of the Elementary and Secondary Education Act (ESEA), a task that the Congress is undertaking during the 1999–2000 session.[1] The proposal is based on powerful evidence that *all* of our children can learn to far more challenging standards than we have hitherto understood, as well as a belief that access to high-quality

[1] At the time of the seminar, I was serving as Under Secretary of Education and Acting Deputy Secretary of Education at the U.S. Department of Education. In February 2000, I moved to Stanford University as a Professor of Education, having previously served on the faculties of Harvard University and the University of Wisconsin and also as Dean of the Graduate School of Education at Stanford University.

education resources is a fundamental right for all children.[2] Our work on the ESEA proposal has been, and will continue to be, informed by efforts to identify and understand interventions that promote both educational excellence and educational equality.

The search for such interventions is not new. During the late 1960s and early 1970s, I was one of a small group of researchers who spent months examining hundreds of computer printout pages, searching for explanations of black-white test score differences. Typically, the data on those printouts were from the "Equality of Educational Opportunity Survey" (Coleman et al. 1966). Those cross-sectional data were not as powerful for our purposes as the longitudinal data now available in the High School and Beyond (HS&B) survey and its successor longitudinal surveys, such as the National Education Longitudinal Study (NELS), or as the trend data now available from the National Assessment of Educational Progress (NAEP). Researchers have obtained a clearer picture of trends in achievement and of the nature of the achievement gap as we have gradually accumulated more national and state NAEP assessment data over time. The clearer picture and the trend data are invaluable in enlightening our efforts; still, we face many challenges in both understanding and reducing the gap. These challenges are, of course, the reasons for convening this seminar.

In this paper, I will first examine some plausible hypotheses for the narrowing of the gap that occurred between 1971 and 1988. Second, I will consider reasons why the gap did not continue reducing after 1988 and will examine NAEP trend data for 1990–1996, with a special focus on 1990. Third, I will propose the argument that reforms in the education system from 1992–1999 in many states are leading, at present, to improvements in teaching and learning and will lead eventually to improvements in student achievement. Moreover, I believe that, if well implemented, these changes will lead to further reductions in the gap. Then I will review 1990–1998 NAEP data from assessments that are more extensive than the NAEP trend assessments. These new data indicate that the reforms may already be supporting overall increases in test scores for both African Americans and whites. Finally, I will comment on the use of experimental methodology in education research and raise some questions about how to measure student achievement.

---

[2] See O'Day and Smith (1993) for a discussion of the impact of standards-based reforms on equality of educational opportunity.

The Achievement Gap Narrows: 1971–1988

In 1990, Jennifer O'Day and I (1991a) examined the achievement results on NAEP trend assessments administered from 1971 through 1988. Over that 17-year time period, the NAEP trend data show that the gap in test scores between African American and white students narrowed substantially for both reading and mathematics achievement. Our analysis considered a number of plausible hypotheses for this reduction in the black-white achievement gap between 1971 and 1988. What we saw—and tried to explain—was a pattern of consistent and substantial increases in African American achievement and almost no change in white scores. The increases in black scores and the relative stability of white scores, taken together, produced a reduction in the gap between the reading scores of black and white students—in less than two decades—of approximately 33 percent for 9-year-olds and over 50 percent for 13- and 17-year-olds. For math, the reductions ranged between 25 and 40 percent over the three age levels. Put in another metric, on the reading assessments administered in 1971 and 1988, the original four grade-level difference between 17-year-old African American and white students in reading narrowed by well over two grade levels.

There was, however, a cloud on this otherwise bright horizon. In our paper, O'Day and I (1991a) suggested that the factors that we believed contributed to the reduction in the gap in the past might have run their course and that other factors were appearing that could lead to a widening of the gap. Two years later, in the context of a paper on educational equality and standards-based reform, O'Day and I (1993) returned to consider the black-white test score gap. To our dismay we found, in the 1990 test score data, preliminary evidence on reading achievement for our hunch. I will address this issue later in this paper.

Now, let us examine the trend data where the reductions in the gap became clear. Table 1 shows the NAEP long-term trend reading scores for whites and blacks at three age levels for test administrations from 1971 to 1996. Table 2 displays the NAEP long-term trend mathematics scores for the same three age levels for test administrations for a shorter period of time, from 1973 to 1996. [3]

---

[3] Tables 1 and 2 in this paper are taken from *NAEP Trends in Academic Progress, 1996*, rather than from Smith and O'Day (1991a).

Table 1. NAEP Long-term Trend Reading Scores, 1971–1996: National, Public and Private School Students

	AGE 9				AGE 13				AGE 17			
Cohort	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff
1954									**1971**	291.4	238.7	52.7
1958					**1971**	260.9	222.4	38.5	**1975**	293.0	240.6	52.4
1962-63	**1971**	214.0	170.1	43.9	**1975**	262.1	225.7	36.4	**1980**	292.8	243.1	49.7
1966-67	**1975**	216.6	181.2	35.4	**1980**	264.4	232.8	31.6	**1984**	295.3	263.6	31.7
1971	**1980**	221.3	189.3	32.0	**1984**	262.5	236.3	26.2	**1988**	294.7	274.4	20.3
1973									**1990**	296.6	267.3	29.3
1975	**1984**	218.2	185.7	32.5	**1988**	261.3	242.9	18.4	**1992**	297.4	260.6	36.8
1977					**1990**	262.3	241.5	20.8	**1994**	295.7	266.2	29.5
1979	**1988**	217.7	188.5	29.2	**1992**	266.4	237.6	28.8	**1996**	294.4	265.4	29.0
1981	**1990**	217.0	181.8	35.2	**1994**	265.1	234.3	30.8				
1983	**1992**	217.9	184.5	33.4	**1996**	267.0	235.6	31.4				
1985	**1994**	218.0	185.4	32.6								
1987	**1996**	219.9	190.0	29.9								

NOTE: This table does not appear in Smith and O'Day (1991a).

SOURCE: *NAEP Trends in Academic Progress, 1996*. National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

Table 2. NAEP Long-term Trend Mathematics Scores, 1973–1996: National, Public and Private School Students

	AGE 9				AGE 13				AGE 17			
Cohort	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff
1961					**1973**	274.0	228.0	46.0	**1978**	305.9	268.4	37.5
1965	**1973**	225.0	190.0	35.0	**1978**	271.6	229.6	42.0	**1982**	303.7	271.8	31.9
1969	**1978**	224.1	192.4	31.7	**1982**	274.4	240.4	34.0	**1986**	307.5	278.6	28.9
1973	**1982**	224.0	194.9	29.1	**1986**	273.6	249.2	24.4	**1990**	309.5	288.5	21.0
1975									**1992**	311.9	285.8	26.1
	1986	226.9	201.6	25.3	**1990**	276.3	249.1	27.2	**1994**	312.3	285.5	26.8
1979					**1992**	278.9	250.2	28.7	**1996**	313.4	286.4	27.0
1981	**1990**	235.2	208.4	26.8	**1994**	280.8	251.5	29.3				
1983	**1992**	235.1	208.0	27.1	**1996**	281.2	252.1	29.1				
1985	**1994**	236.8	212.1	24.7								
1987	**1996**	236.9	211.6	25.3								

NOTE: This table does not appear in Smith and O'Day (1991a).

SOURCE: *NAEP Trends in Academic Progress, 1996*. National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

For both tables 1 and 2, a step-wise dashed line separates the data from before and after 1990; therefore, the reduction in the achievement gap that occurred between 1971 and 1988 can be revealed more clearly; and the "cloud on the horizon" represented by the 1990 data can be seen as it first appeared. One way of interpreting these tables is to think of 10 scale points as being roughly equal to a grade level. (A more precise estimate is that a grade level of growth between 9- and 13-year-olds is equivalent to about 12 scale points, while between 13- and 17-year-olds a grade level is about 8–10 points.) A brief review of these findings indicates *extraordinary* changes, as noted here:

- In reading, over a period of 17 years from 1971 to 1988, the scores for black students tested at 9 years of age increased by 18 scale points. Scores for black 13-year-olds increased by 20 points. Both increases represent an improvement of roughly 1.5 grade levels. During the same period, the scores for white students increased only slightly. Thus, the gap between black and white scores was reduced by almost 1.5 grade levels. Over the same years, scores for 17-year-old black students increased by substantially more than 3 grade levels, and the gap reduced from 53 to 20 points.
- In math, over a period of only 13 years from 1973 to 1986, the scores for 9-year-old black students gained 1.5 grade levels, and the gap narrowed by 0.5 grade levels. The 13-year-old black students improved by almost 22 points, over 1.7 grade levels, and the gap narrowed by almost 2 grade levels. Black 17-year-olds increased their scores by approximately 1 grade level, and the gap narrowed by 1.5 grade levels.[4]

Another way of looking at these data is to consider what happens to a group, or birth cohort, of students who are born in a certain year, in other words, to understand the context of their schooling. Fortunately, the NAEP trend data give us the opportunity to review data for several birth cohorts of students. For example, we have data for students born in 1971 for all three age levels that were administered the NAEP reading assessments: the 9-year-olds took the test in 1980, the 13-year-olds in 1984, and the 17-year-olds in 1988. Looking at these data by cohort gives us the chance to examine whether changes in the test scores and in the gap occurred in the early or later years of school-

[4] See also the discussion in Smith and O'Day (1991a).

ing. Table 1 displays reading scores by cohort.[5] Looking above the dashed line, what do we see in table 1 for data collected prior to 1990? Two points seem important. First, following the progress of the cohorts, we see the gap clearly closing. But we also see that the changes did not happen only during the years from 1971 to 1988. The changes stretch from 1962 to 1988. The 9-year-olds tested in 1971 were born in 1962. This reminds us that, when we explore possible ideas about the causes of changes in scores, we need to think about the time periods the students lived through as they were growing up.

Second, these data on reading achievement give us insights about the effects of schooling on students. The data for the youngest group, the 9-year-olds, may be viewed as reflecting a combination of student background factors (family status, preschool attendance, etc.) and the early years of schooling (e.g., grades K–4). The reading scores of 9-year-old black students from the 1962 cohort to the 1971 cohort show an increase of more than 1.5 grade levels. Recall that these 9-year-olds were tested at the beginning and the end of the 1970s, a period that included the maturing of both Head Start and Title I. The scores for black 9-year-olds in the latter two cohorts (1975 and 1979) showed no increases. White 9-year-olds showed a slow increase over the five cohorts. This resulted in a reduction in the size of the gap, suggesting that, during the early years of this period of time, there were substantial relative improvements in preschool and early elementary school opportunities for black students. And the test score increases from this early period were sustained.

But what happens during the periods from 9- to 13- and from 13- to 17-years-old across the cohorts? Of particular interest is that black students generally "grew" more than whites during the 9- to 13-year age period on the reading test. In this age period, for each of the cohorts, the growth in reading achievement for blacks (about 53 points on the average) exceeded that of whites (about 45 points on the average) with a striking difference of 14 points in the 1975 cohort. The picture is more complicated when the 13- to 17-year-old growth estimates are examined. One fact that stands out is that, in these years, the growth for both groups is, on average, only about 30 scale points or about 60 percent of the average growth for the 9- to 13-year-olds. Thus, there is greater growth on these tests for young adolescents than for older adolescents.

---

[5] Note that for some of the cohorts we took the liberty of using immediately adjacent years. For example, in the 1962–63 combined cohort, both the 9- and the 13-year-olds who were given the test were born in 1962, while the 17-year-olds were born in 1963.

Another feature of these data is that, in the 1958 and 1962 cohorts, the growth for whites was far greater than the growth for blacks. This changed dramatically to no difference in growth in the 1966 and the 1971 cohorts, perhaps indicating some equalization of opportunity in secondary schools.

Plausible Explanations for These Changes

The essay O'Day and I (1991a) wrote was not the end product of a comprehensive research project, but a more general academic overview of education reform. Rather than a detailed statistical analysis, we set out to write a provocative analysis that reviewed broad trends in education policy over the years and related them to changes in test scores.[6] O'Day and I were writing almost 25 years after *Equality of Educational Opportunity* (Coleman et al. 1966) was published, so we framed our text according to that report. Coleman and his colleagues had separated their "explanatory" variables into four clusters: (1) *background and home environment*; (2) *school context* including the social class and racial composition of the school; (3) *teacher characteristics*; and (4) *school resources* including curriculum, libraries, and other items. For our analysis, we combined the third and fourth clusters into one large school resource category, leaving us with three clusters.

Our approach was simple. We reviewed the status of and trends in a number of key explanatory variables within the three clusters and then looked at the scores in the NAEP trend data; we then attempted to relate the two sets of trends to each other. Naturally we were interested in explanatory variables that showed a relationship to student achievement and had a track record in the research literature. We were particularly interested in variables that showed a positive trend or change for blacks, positive in the sense that they indicated that black achievement might rise due to changes in the environment or conditions that were measured by the variable. We wanted to explain the substantial increases in black achievement that the NAEP trend data revealed. To refine our search even more, we also were interested in the same variables showing only a neutral or negative change for whites, because we wanted to be able to understand why white achievement did not increase. Here I will simply summarize a few of our central conclusions.

---

[6] Seven years later David Grissmer and his colleagues developed what now stands as the definitive quantitative work on NAEP data and the test score gap (Grissmer, Flanagan, and Williamson 1998; Grissmer et al. 1998; Grissmer and Flanagan 1998).

Background Factors

Consider again the period of time from 1962 to 1988. This era spans the period from the beginning of the "Great Society" to the end of the Reagan Presidency. The early years of this period, in particular, witnessed substantial strides in economic well-being for families of African American children, with a smaller improvement for white students. The percentage of black children living in families below the poverty line fell from 65 percent in 1960 to 42 percent in 1980 (Smith and O'Day 1991a). Poverty rates also diminished for families of white children; but a substantially smaller percentage of whites were affected, at least in part because the white percentage in poverty was (and still is) much lower. Given the consistent relationship between poverty and achievement, it is reasonable to identify these changes in economic conditions as contributing to improvements in the achievement of both groups, but to a larger extent for black students.

Another related development occurred between 1970 and 1988. The percentage of mothers of black elementary school children who had completed 12 or more years of schooling nearly doubled, moving from 36 percent to 69 percent, while changing only slightly for white mothers (Smith and O'Day 1991a). Studies of academic achievement consistently find the educational attainment of mothers to have a clear and strong relationship to their children's educational achievement.

Finally, preschool attendance is another factor that is moderately related to student achievement, particularly in the early grades. From 1960 to 1980, preschool attendance increased substantially for low-income children (Smith and O'Day 1991a). Because there was a higher proportion of low-income black families, the increase in preschool attendance gave proportionally greater educational opportunities to black children.

Yet the demographic picture was not entirely positive. During the same period, a few other factors related to poverty and to achievement increased substantially more for blacks than for whites, such as the percentage of single-parent families. Other factors held level for blacks while declining for whites, such as the number of low-birthweight babies (Smith and O'Day 1991a).

Social Context

Two important factors that influenced African American children positively from 1962 to 1988 were rural-to-urban migration and desegregation.

From 1960 through the early 1980s, large numbers of African Americans from the rural South migrated to urban communities in the South and the North. For those who remained in the South, the rapid development of metropolitan areas and widespread economic recovery during the 1970s and 1980s were significant forces of change that created important opportunities related to improved academic achievement (Smith and O'Day 1991a).

Second, from 1969–1972 most Southern states experienced large-scale desegregation of their schools. Desegregation had powerful effects on the racial and social class composition of schools, improving conditions that are generally positively related to student achievement for blacks and whites, though more so for low-income students. And desegregation affected more than just the composition of the schools. In the South, particularly, there were dramatic improvements in the characteristics of the new schools that black students attended, as they moved from segregated all-black schools into newly desegregated, formerly all-white schools. In sum, educational opportunities increased, especially for black students in the South (Schofield 1991).

Table 3 reflects some of these changes. It shows that the reading gains for blacks across the NAEP assessments from 1971 to 1988 were much stronger in the South and border states than in the Northeast and Midwest states, reflecting the forces of rural-to-urban migration, desegregation, and economic recovery. The scores for white students in the South and border states also increased more than in the Northeast and Midwest, especially for 9- and 13-year olds, but their gains were nowhere near as large as the gains for blacks; and consequently, the gap shrunk by a large amount.

Again, however, the picture was not completely positive. During the late 1970s, the 1980s, and into the 1990s, concentrations of urban poverty and the racial isolation of African American students increased especially in the North. Considerable research indicates that these factors have a negative effect on student achievement (Smith and O'Day 1991a).

School Characteristics

In our analysis, O'Day and I (1991a) considered a variety of school factors. Two factors, beyond the effects of desegregation in the South, deserve mention in this review. The first was an increased focus across the country on supplemental or compensatory education for children from low-income families between 1966 and the early 1980s. The policy emphasis started with the

Table 3. NAEP Long-term Trend Reading Assessments 1971–1988: North versus South, Public School Students Only

Northeast and Midwest States

	AGE 9						AGE 13						AGE 17					
Birth Cohort	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE
1954													**1971**	295.3	248.4	46.9	1.02	0.19
1958							**1971**	265.9	229.8	36.2	1.01	0.07	**1975**	296.0	246.2	49.8	1.13	0.14
1963	**1971**	219.5	179.3	40.2	0.96	0.12	**1975**	265.0	230.1	34.9	0.98	0.06	**1980**	290.9	247.4	43.5	1.04	0.21
1967	**1975**	220.2	186.4	33.9	0.88	0.05	**1980**	265.5	233.0	32.5	0.93	0.06	**1984**	292.7	264.6	28.1	0.70	0.06
1971	**1980**	222.2	192.4	29.8	0.79	0.08	**1984**	260.5	238.2	22.3	0.63	0.07	**1988**	295.7	273.8	22.0	0.59	0.09
1975	**1984**	218.6	189.5	29.1	0.71	0.07	**1988**	260.7	240.4	20.3	0.58	0.10						
1979	**1988**	221.1	187.7	33.5	0.81	0.08												
1971–1988		**1.6**	**8.4**		**-0.14**			**-5.3**	**10.6**		**-0.43**			**0.4**		**25.4**		**-0.43**

Note: Northeast and Midwest states include CT, MA, NJ, PA, RI, VT, NH, ME, IL, KA, MI, MN, NE, OH, WI, IA, ND, and SD. South and Border states include AL, AR, FL, GA, LA, MS, NC, SC, TN, TX, VA, DE, DC, KY, MD, OK, WV, and MO.

NOTE: This table does not appear in Smith and O'Day (1991a).

SOURCE: Special tabulations, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

Table 3. NAEP Long-term Trend Reading Assessments 1971–1988: North versus South (continued)

South and Border States

	AGE 9						AGE 13						AGE 17					
Birth Cohort	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE
1954													**1971**	290.4	237.2	53.2	1.16	0.08
1958							**1971**	256.4	218.8	37.7	1.05	0.06	**1975**	289.6	238.9	50.7	1.15	0.12
1963	**1971**	207.0	164.2	42.9	1.02	0.08	**1975**	260.0	224.8	35.3	0.98	0.06	**1980**	291.9	239.8	52.1	1.25	0.26
1967	**1975**	211.7	178.2	33.4	0.87	0.06	**1980**	260.9	229.3	31.6	0.91	0.13	**1984**	293.3	261.9	31.4	0.78	0.05
1971	**1980**	218.1	186.3	31.7	0.84	0.10	**1984**	262.0	232.5	29.6	0.83	0.07	**1988**	292.4	271.4	21.0	0.57	0.09
1975	**1984**	213.0	183.7	29.3	0.71	0.07	**1988**	259.7	243.3	16.4	0.47	0.13						
1979	**1988**	217.3	188.9	28.4	0.69	0.09												
1971–1988		**10.3**	**24.7**		**-0.33**			**3.2**	**24.5**		**-0.58**			**2.0**	**34.2**		**-0.60**	

Note: Northeast and Midwest states include CT, MA, NJ, PA, RI, VT, NH, ME, IL, KA, MI, MN, NE, OH, WI, IA, ND, and SD. South and Border states include AL, AR, FL, GA, LA, MS, NC, SC, TN, TX, VA, DE, DC, KY, MD, OK, WV, and MO.

NOTE: This table does not appear in Smith and O'Day (1991a).

SOURCE: Special tabulations, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

Great Society and eventually characterized much of the federal education effort. Title I of the Elementary and Secondary Education Act was the cornerstone program of this effort, providing supplemental assistance to millions of students. Title I programs provided substantial resources and focused attention on the need, particularly in high-poverty schools, to provide low-achieving students with the kinds of support that they needed to learn the "basics."

It is difficult to disentangle the effects of Title I from the effects of other programs, primarily because the program served the universe of high-poverty elementary schools. There are no data sets that show large, specific effects for the Title I program during this period; nevertheless, the program put substantial supplemental educational resources into schools, helped in many districts to lower class sizes, and emphasized the basics of reading and mathematics (Kaestle and Smith, 1982). Perhaps more important, Title I served as a national stimulus and symbol to focus attention on the needs of low-income students. Many Title I students were African American. In my view, it was no coincidence that the growth and expansion of Title I and of "the national idea" of supplemental education support came at the same time as substantial increases in African American scores on NAEP reading and mathematics assessments at the 9- and 13-year-old levels.

These changes in curriculum and in schooling were largely confined to the elementary and middle school levels; however, substantial decreases in the achievement gap between black and white students also occurred at the secondary level from the mid-1970s through the late 1980s. O'Day and I (1991a) argued that this happened, in part, because many secondary schools were also focused on compensatory education. The stimulus in this case came from minimum competency exams. By the mid-1980s, 33 states required passage of a minimum competency test of basic skills as a criterion for graduation (Office of Technology Assessment 1992). We believed that the use of minimum competency assessments had a powerful ameliorative influence on the achievement gap. The instructional emphasis on basic skills, combined with high-stakes testing, produced a greater degree of focus and coherence in the core curriculum of many secondary schools, which might have been lacking in prior years.

Note that, at both elementary and secondary levels, the school-based policies that appear to have affected the gap were focused on basic rather than advanced skills. This focus is consistent with the effects on achievement results during this time period. In summary, the effects were generally positive

for most students who started the era with relatively low scores and neutral for students who started with higher scores. Further, these results are also consistent with the character of the NAEP trend assessment instrument, which itself measures basic skills primarily rather than more complex intellectual abilities.

The Gap Ceases to Narrow: Beyond 1988

Through the 1988 test administration, the NAEP trend data for reading and mathematics achievement showed a powerful and consistent reduction in the gap between the scores of African American and white students at all three age levels. From observation of these achievement data, there was no reason to believe that this trend would not continue. Yet O'Day and I (1991a) argued that the gap would stop narrowing unless some new policies were quickly put into place.

Our argument then was based on three points. First, each of the variables that we believed contributed to the reduction of the gap in black and white achievement had changed by 1990. For example, poverty was slightly increasing during the late 1980s for black families, rather than decreasing. Rising poverty rates would tend to reduce rather than increase average black achievement. Another example is that desegregation and the migration of black families North and to the cities was over; and the concentration of poverty in the Northern inner cities probably was making a negative, rather than positive, contribution to reducing the gap. Finally, the number of states using minimum competency tests as a graduation requirement had stopped increasing.

Our second point was that during the 1980s the national idea of assisting the poor through government programs was under attack. The focus on *equality* had diminished and had given way to a new focus on *quality*. The 1983 report, *A Nation at Risk*, had revealed the inadequate performance of American students compared to their international peers (National Commission on Excellence in Education 1983). This report stimulated the already growing interest among the states, especially in the South, in reforming the schools through such measures as increasing the number of hours in a school day, adding to the number of days in a school year, and encouraging students to take tougher courses.

While an overall increase in quality would support children from low-income as well as affluent families, two observations indicated that these reforms

would not reduce the gap. One observation is that for years the United States had condoned a curriculum and a quality of instruction in low-income inner-city and rural schools that are shallow and insufficient compared to those available to their suburban and well-to-do peers. There was little in the new *quality* reforms that would redress this *inequality* in opportunity. Another observation is that O'Day and I realized that, even if there were significant improvements in curriculum and instruction, they would take considerable time to implement and, further, that schools in inner cities and poor rural areas lag their counterparts in the suburbs in carrying out such changes. Thus, even though we hoped we were incorrect, we suggested that poor and minority students were less likely to benefit from the new focus on higher quality, at least in the short run.

Finally, the third point was that, while black students had shown striking increases in their scores between 1971 and 1988, there was reason to believe that many of the gains on the assessments that could be achieved from across-the-board emphases on the basics and minimum competencies had already been achieved by the late 1980s. Thus, in the future, closing the gap would require greater attention to higher-order skills such as reading comprehension and problem solving than poorly performing students had received in the past.

On the basis of these conjectures, O'Day and I questioned whether the gap would continue to reduce in size over the decade of the 1990s (Smith and O'Day 1991a). We suggested that it would *not* continue to narrow unless three vigorous steps were taken to differentially improve the quality of educational opportunity for African Americans. In particular, we argued that, in the absence of an effective policy to alleviate poverty, the country needed an aggressive effort to support low-income families with children and to prepare all children for school by improving their access to quality health services and early childhood education.

Outside of schools, we advocated increasing the opportunities for school- and community-based after-school programs for students in inner-city and poor rural areas, in order to provide safe and academically stimulating environments beyond the six-hour school days. Within schools, we argued for rapid movement toward state standards-based reform, a relatively new idea then, though well understood now.[7] Our argument then centered on the need to eliminate the

[7] See Smith and O'Day (1991b) for additional details.

decades-long practice of giving our neediest students in the highest poverty schools the least-trained teachers and a "watered-down" curriculum. The curriculum and the quality of instruction needed to be upgraded throughout many schools, but the need was greatest in high poverty areas. I will return to these policies in my discussion of how to address the achievement gap, in the last section of the paper.

Now, what happened after 1988? In 1992, O'Day and I wrote a second paper in which we specifically examined the results of the 1990 assessment (O'Day and Smith 1993). To understand what happened to the achievement gap in 1990, we need to refer once again to table 1 (p. 252). The most dramatic changes happened in reading achievement. For the 17-year-olds in the 1971 cohort, who were tested in 1988, the test score difference between African American and white students was 20 points, down 32 points compared to 1971. Then, suddenly, in 1990 the gap enlarged to 29 points. For 9-year-olds the gap increased by 6 points, and for 13-year-olds by only 2 points. For mathematics achievement at the 9- and 13-year-old levels there was little change in the gap from 1986 to 1990, while for the 17-year-olds the gap narrowed somewhat. Now, of course, this was only one new point in time, and one data point is clearly insufficient for making strong inferences about changes in a long-term trend. Still, for reading achievement, the 1990 assessment revealed sudden, substantial changes in the size of the achievement gap in precisely the direction that we feared. Further, for two of the age levels in mathematics, the gap remained fairly constant between 1986 and 1990 (see table 2, p. 253).

We now have the luxury of looking over a longer period of time. As a number of authors in Jencks and Phillips (1998) suggest, the NAEP longitudinal trend data from 1990 through 1996 do not show a clear pattern of growth for either African American or white students in either reading or mathematics achievement. Consequently, for these trend data there is no clear pattern of change in the gap during the early and middle 1990s. It seems clear that the gap had ceased to reduce, at least as measured by the NAEP trend data. But the NAEP trend data may not be telling the entire story.

The 1990s: Standards-Based Reform

During the decade of the 1990s, the nation's focus on educational improvement and reform was unprecedented. Federal, state, and local governments

made education their top priority. At the federal level, three fundamental objectives guided investment strategies in K–12 education, as follows:

1. Create economic and health care environments as stable and livable as possible for all families with children.
2. Expand opportunities for all students to participate in engaging and educationally rich activities beyond the traditional school hours. These opportunities include high-quality preschool and after-school opportunities, particularly for children from the least affluent families.
3. Stimulate and support state and local standards-based reform strategies to improve the quality of schools for all students.

For each of the objectives, substantial progress has been made by the Administration, Congress, and the states, though there continues to be considerable distance between current conditions and the fulfillment of these ambitious goals. For the first objective, sustained economic growth throughout the middle and late 1990s created over 10 million new jobs, and the unemployment rate has dropped to record lows. In addition, the Earned Income Tax Credit (EITC) provided millions of low-income families with children additional income to lift them above the official poverty level; and the Children's Health Insurance Plan (CHIP) makes it possible for every child in a low-income family to have adequate health care.

The second objective has seen the development of education standards for the Head Start curriculum and the expansion of Head Start enrollment, which has brought 72,000 children into the program since the reauthorization in 1994.[8] In addition, the Administration has dramatically expanded the 21st Century After-School Program, from a $1 million program in FY1996 to a $450 million program in FY2000 that will serve well over half a million students during the school year 1999-2000. Both of these programs are targeted to provide services to low-income students who would not otherwise have either preschool or after-school educational opportunities.

Finally, the nation's standards-based reform movement is based on the principle of establishing coherent and fair policies at the state and local levels to improve student learning (Smith and O'Day 1991b). Standards-based re-

[8] See www2.acf.dhhs.gov/programs/hsb.

form has taken on a life of its own and is now the dominant reform in most states—it has become a national idea. There are four fundamental elements of this reform:

1. Establish challenging state content and performance standards for all students.
2. Align all parts of the education system to assist all students to learn the content, skills, and strategies set out in the state standards. The curriculum, teacher training, technical assistance, and assessments should all be based on the state content and performance standards.
3. Provide local school districts and schools with sufficient fiscal resources and at the same time grant sufficient flexibility, autonomy, and responsibility to the districts to use their resources to maximize students' opportunities to achieve to the standards.
4. Develop and implement accountability systems that use student performance measures to demonstrate to the public that schools and districts are meeting their obligations to teach all students to challenging content standards.

This reform strategy has shaped federal and state education policies throughout the decade. At the federal level, Goals 2000, a separate program proposed by the Clinton Administration and passed in 1994, provides support to states to develop and implement standards-based reforms. The core federal programs in education (including Title I, the program for supplemental services for low-achieving students) were reauthorized and modified to support the state reforms. Beginning in 1994, Title I legislation required that all students eligible for Title I services be taught to the same challenging state standards as all other children in the state—this requirement focuses specifically on closing the gap. This standards-based reform strategy has provided a focus for a variety of other supportive interventions, including the use of technology in the schools, the professional development of teachers, and even the stimulation of charter schools.[9]

In 1993, few states had any type of coherent content standards. In 1998, 44 states had content standards in at least three subjects (*Education Week* 1998).

---

[9] Secretary of Education Richard Riley, in his State of American Education Address (February 22, 2000), stated the number of charter schools as about 1,700. See www.ed.gov.

Most states have their assessments aligned with these content standards (*Education Week* 1998). In most districts, the curriculum and the textbooks are examined and selected based upon their alignment to the state standards. A substantial number of states are also establishing programs of teacher training and teacher professional development designed to prepare teachers to teach to these new content and performance standards. Though it may take several years to achieve full implementation in many states, I believe that the cumulative force of such coherent strategies and policies in so many states is already having positive effects on student achievement.

New Assessment Data from NAEP: Three Policy Perspectives

To help explore this belief, I want to introduce a new set of NAEP data into the conversation—data from the "main" NAEP. The NAEP trend data presented in the previous sections of this paper came from a supplemental assessment that has been administered since 1971 strictly for the purpose of maintaining longitudinal trends. This NAEP trend assessment has undergone few changes in format or content since 1971. As a consequence it does not measure some of the concepts, knowledge, and skills that have been introduced into the curriculum in more recent years.

Another NAEP assessment—called the main NAEP—serves now as the primary measure of the achievement of the nation's students. Since 1990, the main NAEP has measured a wider range of skills and knowledge with a broader repertoire of item types than the trend NAEP. The 1990 version is administered on a regular schedule at the national level and also on a voluntary basis by many states. The main NAEP data show promising signs of the effectiveness of the state standards-based reform movement. Three views of data from the main NAEP are illustrated in tables 4, 5, and 6 below.

Reading Achievement

Table 4 sets out national results for reading at three grade levels (fourth, eighth, and twelfth) for whites, African Americans, and Hispanic Americans. Reading scores were collected in each of three years: 1992, 1994, and 1998. During the period 1992–1994 the effects are generally negative; with one exception the scores dropped over those years, and the losses were greater for African Americans and Hispanic Americans than for whites.

However, from 1994 to 1998 the reading scores for all groups increased; and the increases were slightly larger for African Americans and Hispanic Americans than for whites. Why is there a change in direction of these scores? One hypothesis is that the 1994 test administration is an anomaly—that it somehow represents a mismeasurement—and that we should therefore overlook the results and simply note the difference between the 1992 and the 1998 results. For grades fourth and eighth, this comparison reveals some slight increases over the 6 years, except for fourth grade Hispanic Americans; and for grade 12, slight decreases.

There is a second possibility. Since Goals 2000 did not pass until 1994 and many states did not begin their reforms until then, one could argue that 1994 is a better baseline year than 1992 for estimating the effects of the new state reforms. The gains from 1994 to 1998 are large for all groups at all grades, suggesting that the reforms may be having a positive effect. Whatever the case, the scores are now moving in the right direction. Note, however, that there is no evidence of the black-white gap's closing.

Mathematics Achievement

The overall picture is more optimistic for mathematics, as seen in table 5. These data show strong gains in achievement for all but one of the comparisons from 1990 to 1992 and from 1992 to 1996. Unfortunately, we do not have 1994 or 1998 data for mathematics achievement. Thus there is no true baseline data point to measure the effects of the reforms. Overall, from 1990 to 1996, the gains are more than 1.0 grade levels for all but three of the nine comparisons in table 5; for those three, they are between 0.5 and 1.0 grade levels. The effects here are difficult to attribute to standards-based reform because of the lack of a good baseline, but they are consistent with the possibility of a positive effect from the reforms.

Thus, the picture for the main NAEP data for the 1990s indicates a potentially positive effect of the state standards-based reforms on both reading and mathematics achievement. These data by no means present an ironclad argument, but they are suggestive. Again, there is no sign of a reduction in the gap.

State Trends

A second perspective views the results from the state NAEP as additional support for the argument concerning the effectiveness of the reforms. David

Table 4. Reading Achievement: NAEP National Data by Year, Grade, and Racial-Ethnic Group

	1992	1994	1998	Difference in Scale Scores	
				1994–1992	1998–1994
4th Grade					
White	225	224	227	-1	+3
African American	193	187	194	-6	+7
Hispanic	201	191	196	-10	+5
8th Grade					
White	267	268	272	+1	+4
African American	238	237	243	-1	+6
Hispanic	241	240	244	-1	+4
12th Grade					
White	298	294	298	-4	+4
African American	273	265	270	-8	+5
Hispanic	278	270	275	-8	+5

NOTE: Includes both private and public school students.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992, 1994, and 1998 Reading Assessments.

Table 5. Mathematics Achievement: NAEP National Data by Year, Grade, and Racial-Ethnic Group

	1990	1992	1996	Difference in Scale Scores	
				1992–1990	1996–1992
4th Grade					
White	220	228	232	+8	+4
African American	189	193	200	+4	+7
Hispanic	198	202	206	+4	+4
8th Grade					
White	270	278	282	+8	+4
African American	238	238	243	0	+5
Hispanic	244	247	251	+3	+4
12th Grade					
White	301	306	311	+5	+5
African American	268	276	280	+8	+4
Hispanic	276	284	287	+8	+3

NOTE: Includes both private and public school students.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1990, 1992, and 1996 Math Assessments.

Grissmer of RAND was asked by the National Goals Panel to look behind the results of state NAEP and to try to explain why North Carolina and Texas performed better compared to other states during the 1990s (Grissmer and Flanagan 1998). Grissmer and his colleague, Ann Flanagan, found, among other things, that both North Carolina and Texas have implemented standards-based reforms in several subject areas over the last few years. These states maintain consistent policies that emphasize relatively challenging standards, require curriculum-aligned tests, provide for accountability at the school level, offer extensive teacher training, and focus special efforts on low-scoring students. In addition to the focus on teaching and learning, Grissmer and Flanagan found that the financial support and committed involvement of the business community and the sustained focus on education in government, despite partisan shifts in the political leadership, were also positive influences on the effectiveness of the reforms.

To illustrate the effects of such coherent policies, table 6 presents state NAEP data on fourth graders from the 1992 and 1996 mathematics assessments at two achievement performance levels, making possible comparisons between state-level and nationwide data. Nationwide, the percentage of white students who scored at or above the Basic (the first level) performance level increased from 69 percent in 1992 to 74 percent in 1996. For blacks nationwide, the percentage of students scoring at or above Basic increased from 22 percent in 1992 to 32 percent in 1996. For Hispanic American students nationwide, the increase in students scoring at or above Basic was from 33 percent in 1992 to 40 percent in 1996. In other words, each group improved; and, on this measure, minority fourth graders across the country may be once again beginning to close the achievement gap, though there is still a long way to go.

Table 6 shows clearly the difference between the results in Texas and North Carolina compared to the national results and also to student performance in three other states that serve as rough benchmarks—California, Florida, and New York. In Texas, which has focused intense efforts on improving performance in low-scoring schools, white fourth graders scoring at or above Basic moved up from 72 percent in 1992 to 85 percent in 1996. For blacks, the increase at or above Basic was from 29 percent to 47 percent. For Hispanic American students, the percentage increase was from 43 percent to 55 percent. The data for North Carolina reveal similar increases for both black and white students. For whites, the percentage achieving at or above the Basic level went

Table 6. 1992 and 1996 Mathematics National Assessment, Percentage of Students at *At Or Above Basic* Achievement Level by Race-Ethnicity, Grade 4 Public School Students

	White		Black		Hispanic	
	1992	1996	1992	1996	1992	1996
Texas	72	85	29	47	43	55
North Carolina	65	77	24	37	35	43
California	61	63	21	18	27	29
Florida	66	70	22	26	27	29
New York	71	80	31	37	33	40
Nation	**69**	**74**	**22**	**32**	**33**	**40**

SOURCE: These scores are from the main NAEP assessments, not from long-term NAEP trend assessments.

from 65 in 1992 to 77 in 1996; for blacks, the percent scoring at or above Basic increased from 24 in 1992 to 37 in 1996.

In contrast to North Carolina and Texas, California had a comparatively incoherent and weak strategy for school reform during the early and middle years of the 1990s, and the numbers in table 6 seem to track that incoherence. For white students, the percentage achieving at or above the Basic level in California barely moved upward from 61 in 1992 to 63 in 1996. For black students, the percentage dropped from 21 to 18, and for Hispanic students, the percentage moved up a small amount, from 27 to 29. A comparison of the data between California and Texas (which is somewhat similar to California in the size and diversity of its student population) underscores the fact that, for one reason or another, some states did well in their efforts to improve student outcomes, while others did not.

State Reform Rankings and State NAEP Data

A third and admittedly speculative perspective on examining the effects of the reforms uses the state NAEP data for reading for 1994 and 1998. The trade newspaper, *Education Week*, published an analysis of the progress of the various states on standards-based reform as a stand-alone insert to its publication (*Education Week* 1998). As part of its analysis, *Education Week* ranked the quality of the state reforms according to a number of criteria that represented whether they had achieved certain components of the reform up to that time.

I hypothesized that, if the rankings were valid and the reforms were having an effect on achievement, then the ranking of the reforms should correlate with the gains that the states made on NAEP. I used fourth grade reading gains by state over the period 1994–1998. Because only 33 states had administered the state NAEP in both 1994 and 1998, the sample was somewhat constrained. Nonetheless, after controlling for differences in state per pupil expenditures, the partial correlation between the rankings and the gain scores was +0.43; after controlling for state poverty levels, the partial correlation was +0.46. In each case the partial correlation was statistically significant at the .05 level on a one-tailed test of significance.

Final Remarks: Research Priorities and Methodological Issues

A major purpose of this conference is to examine methodological issues and their inter-relationship with the capability of education research to understand the perplexing questions regarding the test score gap. The importance of these issues is intensified because the dynamics of schooling, and possibly the causes of the achievement gap, are changing. The extraordinary developments in information technology, a new and demanding market economy, the concentration of sustained poverty among a large percentage of families with children (particularly African American and Hispanic American families), and a wide variety of immigrants speaking many different languages are transforming both our expectations and our requirements for schooling.

In this context, I would like to explore themes in two areas. The first theme is the use of experiments to provide more reliable and valid research information about how to help close the achievement gap. The second area has to do with measurement concerns that emerge in thinking about the achievement gap, such as how we measure performance against a standard. Making progress in each of these areas is important to creating a strong foundation of evidence upon which solid policy can be built.

The Use of Experiments

Two general points are important when evaluating research. First, a good model grounded in theory should facilitate explanations of results, and carefully designed measures should provide clear understanding of the data in any study. The second point is that every methodology has weaknesses and strengths

and that, further, it is important to understand these weaknesses and strengths in order to select either a single methodology or a combination of different methodologies to examine the research question. The use of true experiments should be a significant part of the repertoire of federal evaluation programs. Random assignment and deliberate intervention, taken together, provide a useful tool to test hypotheses and estimate effects, whether a theory is strong or weak. When they are appropriate, and if carefully designed and implemented, controlled field experiments can often demonstrate powerful and persuasive evidence.

The potential strength and authority of experimental field trials has been demonstrated by two sets of studies. The first is the set of Tennessee STAR (Student Teacher Achievement Ratio) studies (Finn and Achilles 1999; Word, Johnston, and Bain 1990). The Tennessee class size reduction study in the early grades, initiated over a decade ago, yielded findings that have had an enormous impact on policy debates across the nation. In addition to recent federal legislation (that funds the hiring of additional teachers), several states and local districts are moving toward smaller classes, particularly in the early grades. Nearly all of these legislative initiatives are motivated, at least in part, by the findings of the Tennessee experiments. The fact that this study included randomization of students to classes of different sizes contributed greatly to its influence on policy discussions. Its acceptance was also enhanced by the large number of schools involved in the experiment and by the commonsense nature of the treatment (small class size in the early grades to improve student achievement). Further, the reputation and importance of the STAR study were enhanced by follow-up studies that showed that the initial positive results were sustained in the middle and high school years (Nye et al. 1994).

Another example of the use of experimental methodology can be found in the series of studies on early reading acquisition conducted by the National Institute of Child Health and Human Development (NICHD) under the direction of Reid Lyon (Moats and Lyon 1997; Snow, Burns, and Griffin 1998). These studies consisted of interlocking experiments, conducted over a 10-year period, that examined a set of theoretical hypotheses related to word recognition and reading in young children. The experiments were developed in such a manner that the findings from one study were linked to related hypotheses in other studies. Just as with the STAR study, the fact that the studies were experiments added to their credibility with Congress. The reputation of the NICHD as a more “scientific” agency, rightly or wrongly,

was another factor in the credibility accorded this work. Credibility is often a necessary first step, but it is not enough to ensure that the findings are utilized. The NICHD researchers buttressed their credibility with a commonsense approach and detailed descriptions of their findings in a coherent and compelling fashion. Their story is about how their related set of experiments provided quality insights that can improve the chances of many young children to learn to read. These two examples demonstrate both the feasibility of field experiments and their potential influence in the policymaking process.

As we consider a research agenda for OERI and NCES in the future, several straightforward recommendations from this research seminar for experimental studies could be productive in the clarification of programmatic interventions that can narrow the achievement gap. For example, in the area of promoting better quality in the preparation of students for school, one possibility would be to conduct experiments that address different approaches for training parents to use new understandings of the development of cognitive processes to help their own children. Within schools, the NICHD studies could be extended to cover large and more diverse settings to determine the strength of the reading interventions in less controlled environments. Along the same line, a thoughtful set of experiments that explore the implications of the NICHD research for Limited English Proficient (LEP) children could be important. In an emerging area, literally dozens of potentially powerful field experiments need to be conducted on some of the promising new ways of using technology in classrooms and for distance learning. In this arena, we will need to learn how to carry out experiments in the "real time" necessitated by the rapid changes in the nature of technology. Finally we need experimental data on the effects on students of summer school and after-school programs. Early experiments in the 1970s found few effects for summer schools—these questions need to be revisited as summer schools become more and more a method of expanding students' opportunities to learn.

Methodological Issues in the Measurement of the Gap

Finally, let me suggest two important methodological issues in the measurement of the gap in performance between white and minority students. First, OERI and NCES need to develop a research agenda aimed at understanding what the two types of NAEP assessments (trend and main) *actually* measure. In other words, current efforts to assess student achievement would benefit

from carefully conducted construct validity studies. The data from the two assessments reveal quite different results during the 1990s—it would be nice to know why.

Second, at the present time, various states are struggling toward a set of more sophisticated assessments. These assessments should be aligned with state content standards that set out the skills and knowledge that students should know and be able to do. Also, the assessments should be designed with performance standards or levels that indicate how well a student has learned the skills and knowledge. For example, where a content standard might specify that a student should be able to write a short persuasive essay, performance standards would provide a way of measuring the quality of the persuasive essay. If done well, performance standards would not be established by selecting cut scores on norm-referenced assessments as the NAEP achievement levels are now defined. As assessments improve, performance standards should become real in some sense. For example, reading at or above the Basic level could be validated to show that a student can read and comprehend a well-defined set of books and passages. Similarly, more demanding performance items could define and assess other more challenging levels beyond Basic.

Well-constructed content and performance standards would be the products of the intersection of reasonable theories of the content area, human development, human learning, and pedagogy. Growth through performance levels might be discontinuous rather than smooth. For example, achieving a Basic level of reading in the fourth grade may require effective word attack skills, while achieving a higher level may require mastery of strategies of comprehension. That is, a different dimension of mastery may be required to achieve a higher level, one involving a qualitatively different set of skills and knowledge. This would have substantial implications for the curriculum, as well as for the interpretation of group differences. In these circumstances, the achievement gap would no longer be measured by a scale score difference, but by differences in the percentages of students achieving to the different performance levels. Measurement of the gap becomes more complicated conceptually and methodologically because there are multiple comparisons related to the different performance levels. The problems are already apparent when the trends from the new NAEP assessments and the results of some state assessments are interpreted.

Until we have content and performance standards established on the basis of reasonable theory, we will continue to have political and educational problems with how to set them. A current political issue has to do with how challenging is "challenging," or how do we determine where to set performance levels so that they positively rather than negatively motivate students and educators to succeed? A goal considered impossible to attain by students, teachers, and parents may undermine the credibility of the reforms. Moreover, the size of the achievement gap, as measured using a performance level, may be determined in part by how high the bar is set, that is, by the degree of difficulty of the performance standard. A bar or performance standard that is set relatively low will have higher percentages of students passing and, therefore, will typically result in a smaller gap between groups that have different levels of command of the tested content and skills. Conversely, a bar set very high may make it practically impossible to have high percentages of students pass and may exaggerate the magnitude of differences among groups. This is not an academic issue. In the past the tendency has been to lower the bar to minimize failure. This approach, however, fails to meet the purpose of the reforms to challenge all students to meet rigorous standards. But, now, there also are a number of key states, including New York and Massachusetts, where thoughtful analysts believe that the bar has initially been set too high, and this policy may jeopardize support for the reforms.

Conclusion

Educational excellence and educational equality—again I emphasize that these are the nation's co-equal guiding goals. These are the overarching ideals that researchers and policymakers must continuously keep in mind in their consideration of such important issues as which interventions most effectively enhance student learning, which state systemic structures support achievement in a cost-effective manner, and which collaborations and partnerships can achieve strong public and parental support, while advancing overall improvement in schools in the long term. My hope is that, in the deliberations of this seminar, we will find that we have begun to chart a course of recommended research directions and policy alternatives that will, both sooner and later, assist us in solving the remaining challenges, increase our commitment to the goals, and eventually help in the attainment of those national ideals.

References

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Council of Chief State School Officers. (1994, May). *Baselines for Goals 2000 Implementation.* Washington, DC: Author.

Education Week 18(17). (1998, January 11). Quality Counts. Bethesda, MD: Author.

Finn, J. D., and Achilles, C. M. (1999, summer). Tennessee's Class Size Study: Findings, Implications and Misconceptions. *Educational Evaluation and Policy Analysis* 21 (2).

Grissmer, D. W., Kirby, S. N., Berends, M., and Williamson, S. (1994). *Student Achievement and the Changing American Family*. Santa Monica, CA: RAND.

Grissmer, D. W., Williamson, S., Kirby, S. N., and Berends, M. (1998). Exploring the Rapid Rise in Black Achievement Scores in the United States, 1970–1990. In U. Neisser (Ed.), *The Rising Curve: Long-term Changes in IQ and Related Measures.* Washington, DC: American Psychological Association.

Grissmer, D. W., Flanagan, A., and Williamson, S. (1998). Why Did Black Test Scores Rise Rapidly in the 1970s and 1980s? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.

Grissmer, D., and Flanagan, A. (1998, November). *Exploring Rapid Achievement Gains in North Carolina and Texas.* Washington, DC: National Education Goals Panel. (www.negp.gov/Publications/LessonsfromtheStates).

Jencks, C., and Phillips, M. (1998). (Eds.) *The Black-White Test Score Gap*. Washington, DC: The Brookings Institution Press.

Kaestle, C. F., and Smith, M. S. (1982, November). The Federal Role in Elementary and Secondary Education, 1940–1980. *Harvard Educational Review 52*(4): 384–408.

Moats, L. C., and Lyon, R. (1997, November–December). Critical Conceptual and Methodological Considerations in Reading Intervention Research. *Journal of Learning Disabilities 30*(6): 578–588.

National Commission on Excellence in Education. (1983). *A Nation at Risk: The Imperatives for Educational Reform*. Washington, DC. Author.

Nye, B., Zaharias, J., Fulton, B., Achilles, C. M., and Cain, D. (1994). *The Lasting Benefits Study: A Continuing Analysis of the Effect of Small Class Size in Kindergarten through Third Grade on Student Achievement Test Scores in Subsequent Grade Levels.* Unpublished paper. Tennessee State University, Nashville.

Office of Technology Assessment. (1992). *Testing in American Schools: Asking the Right Questions,* p. 15. Washington, DC: U.S. Government Printing Office.

Schofield, J. W. (1991). School Desegregation and Intergroup Relations: A Review of the Literature. In G. Grant (Ed.), *Review of Research in Education, Vol. 17* (pp. 335–409). Washington, DC: American Educational Research Association.

Snow, C. E., Burns, M. S., and Griffin, P. (Eds.) (1998). *Preventing Reading Difficulties in Young Children*. Committee on the Prevention of Reading Difficulties in Young Children, National Research Council. Washington, DC: National Academy Press.

Smith, M. S., and O'Day, J. A. (1991a). Educational Equality: 1966 and Now. In D. Verstegen and J. Ward (Eds.), *Spheres of Justice in Education: The 1990 American Education Finance Association Yearbook.* New York: Harper Business.

Smith, M. S., and O'Day, J. A. (1991b). Systemic School Reform. In S. Fuhrman and B. Malem (Eds.), *The Politics of Curriculum and Testing: The 1990 Yearbook of the Politics of Education Association* (pp. 223–276). London: Falmer Press.

Smith, M. S., and O'Day, J. A. (1993). Systemic Reform and Educational Opportunity. In S. Fuhrman (Ed.), *Designing Coherent Education Policy* (pp. 250–312). San Francisco, CA: Jossey-Bass Publishers.

Word, E., Johnston, J., and Bain, H. (1990). *Student Teacher Achievement Ratio (STAR): Tennessee's K–3 Class Size Study: Final Summary Report 1985–1990.* Nashville: Tennessee Department of Education.

Educational Research and Educational Policy: An Historical Perspective

Christopher Jencks
Malcolm Wiener Center for Social Policy
John F. Kennedy School of Government
Harvard University

When I began doing quantitative research on education in the mid-1960s, most of what social scientists thought they knew about the effects of educational policies was based on experiments. There were not many experiments, and they did not all reach consistent conclusions, but they were about all we had to go on.[1] School administrators and teachers paid this research very little heed. When they thought about the effects of different educational policies, they based their judgments on personal experience. Educators who had spent years in the schools mostly had strong views about what worked and what did not. Of course, educators often disagreed with one another about what lessons experience taught, but these disagreements seldom led to self-doubt. Nor did educators who disagreed seek to resolve their disagreements by reading educational research.

Although educators seldom sought researchers' advice, researchers continued to offer it. They ran small experiments that tried to assess the accuracy of educators' beliefs about such matters as class size, ability grouping, and the best way to teach reading. More often than not, researchers interpreted their findings as showing that educators' beliefs were wrong. As a result, educational researchers tended to feel superior to those who staffed the schools.

The 1960s: Misinterpreting Insignificant Coefficients

Although educational researchers often felt superior, educators seldom felt inferior. Faced with a study showing, let us say, that children's spelling

---

[1] The opinions expressed in this paper are those of the author. Address any comments to Professor Christopher Jencks, John F. Kennedy School of Government, Malcolm Wiener Center for Social Policy, Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138.

skills did not improve over the long run if teachers handed out a list of words every Monday and gave a spelling test every Friday, most educators simply dismissed the findings as implausible. Likewise, when studies seemed to show that children learned no more in small classes than in large classes, few educators considered the possibility that the studies might be right. In most cases educators did not even bother to dismiss such results as implausible. They simply ignored educational research entirely. Legislators did the same thing.

Educators' indifference to research results convinced many researchers that the people who staffed the nation's schools were unscientific traditionalists, unwilling to consider the possibility that their prejudices were ill founded. In retrospect, however, the educators' indifference to educational research seems largely justified.

Then as now, educational experiments typically assessed policies by comparing outcomes for students who had had different educational experiences. Sometimes these comparisons involved small experiments in which students were randomly assigned to different treatments. Sometimes they involved "natural experiments," in which students had different educational experiences because school boards, principals, or teachers followed different practices. Since there were no national or statewide testing programs that allowed researchers to link students' achievement to past experiences, most of these comparisons were based on small samples. As a result, the sampling errors of the estimates were usually quite large, and the difference between those who had had different experiences was often less than twice its sampling error. Researchers almost always interpreted this finding as supporting the "null hypothesis," namely that the experience in question made no difference.

With the wisdom of hindsight, this interpretation of insignificant coefficients looks foolish (though it is still disturbingly common). Every first-year statistics student learns that data analysts can make two different sorts of errors. "Type One" errors occur when the analyst accepts a false hypothesis as true. "Type Two" errors occur when the analyst rejects a true hypothesis as false. Social scientists have traditionally been far more concerned about avoiding Type One than Type Two errors. This bias makes sense when social scientists are testing their own theories. It may also make sense when social scientists are testing theories of interest only to other social scientists, since most such theories are too simple to be useful and reducing intellectual clutter is always a high priority in science.

In the policy arena, however, most data analysts are testing hypotheses that are widely accepted in the real world, not hypotheses that come from some theorist's fevered imagination. This reality means that when samples are small and measurement imprecise, policy researchers who emphasize significance tests are far more likely to reject a true hypothesis than to accept a false hypothesis. If practitioners were to take such researchers' conclusions seriously, they would often be led badly astray.

With the wisdom born of hindsight, one can see that policy researchers had several better alternatives. Bayesian theory suggests, for example, that researchers should start out by formulating "priors" that describe their best guesses about how the world works. As they accumulate additional evidence, either from experiments or other sources, they should update their priors to incorporate this new information. Had educational researchers tried to proceed in this fashion, their priors about policy questions would presumably have been shaped by two considerations, as follows:

1. If most educators think that a policy enhances student achievement, and if there is no other evidence about the policy's impact, a reasonable person should assume that practitioners are somewhat more likely to be right than wrong.
2. If an educational policy had *very large* effects, this would be obvious to everyone, and we would not be doing research on the policy's impact. Thus, if we are doing research on a policy's effect, the effect is not only likely to be positive, but also likely to be relatively modest.

If researchers had reasoned in this way, they would hardly ever have started out with the null hypothesis—the theory that a popular policy has no effect whatever. Thus, when they generated new data showing that a policy's impact was quite uncertain, they would not have raised the possibility that the policy had no effect. Instead, they would have concentrated on estimating the actual size of the effect.

For educational researchers who had no prior expectations about how large an impact a policy was likely to have, traditional statistical methods offered another attractive option. Researchers could just have reported the odds that a policy had a positive rather than a negative effect. Suppose, for example, that an investigator had randomly assigned first graders to one of two reading classes: a class of 15 and a class of 25. Suppose, too, that at the end of the year

the children in the smaller class scored 0.3 standard deviations higher on the investigator's reading test than the children in the larger class, but the sampling error of this difference was also 0.3 standard deviations. The 95 percent confidence interval for the effect of being in a reading class of 15 rather than 25 therefore runs from –0.3 standard deviations to +0.9 standard deviations.

The most frequent interpretation of such a result is that since the confidence interval includes 0, we cannot reject the hypothesis that class size has no effect on reading achievement. A more plausible conclusion, I would argue, is that since five-sixths of values in the confidence interval are positive, the odds are 5 to 1 that small classes raise reading achievement rather than lowering it. If researchers have strong priors, of course, the odds that small classes raise reading achievement are even higher. The odds that class size has an impact of exactly 0 are, in contrast, vanishingly small.

Thirty years ago, however, researchers almost always emphasized statistical significance when formulating their conclusions. Even today, educational researchers who get statistically insignificant coefficients are more likely to suggest that the variable in question has no effect than to conclude that their sample was too small to justify any firm conclusions. Under these circumstances, I think that educators' skepticism about educational research was largely justified. Unfortunately, educators seldom had enough statistical expertise to explain their skepticism in a technically compelling way. So they just ignored educational research or dismissed it as irrelevant to classroom practice.

The Coleman Report

Educators' refusal to take educational research seriously faced its first important challenge in 1966, when the U.S. Office of Education released a report by James Coleman et al. analyzing the determinants of student achievement in some 4,000 schools throughout the country. Coleman and his colleagues found a weak relationship between many popular educational policies and student achievement, which was nothing new. But Coleman's report differed from earlier studies in several crucial respects. First, Coleman was a distinguished sociologist, whom the Office of Education had selected to carry out a Congressionally mandated study. Second, his analyses covered far more schools and students than any earlier study, so the results could not be dismissed as a fluke. Third, Coleman's work appeared at a time when the federal government was beginning to play an expanded role in educational agenda-setting and when

policymakers were more attentive to the findings of social science than they had been in earlier periods. Fourth, when other social scientists reanalyzed Coleman's data, as many did over the ensuing years, they often faulted his methods, but usually came to broadly similar substantive conclusions.[2]

If educational researchers had been committed Bayesians, they might once again have said: "Well, the estimated effects of educators' preferred policies may be statistically insignificant, but the standard errors of the estimates are so large that we should not draw any strong policy conclusions from them." But we did not say that. I myself once tried to go down that road, and it was a dead end.

My first paper on the relationship between school policy and student achievement was a reanalysis of the Coleman data (Jencks 1971). In it, I reported the confidence interval for each estimated effect. As far as I know, nobody read this paper except the editors of the volume in which it appeared. Certainly no educational researcher concluded that this was a good way to present statistical findings. Nobody wants to read a paper reporting that many different policies could have fairly sizable positive effects, no effect, or a modest negative effect. Papers claiming that popular policies have "insignificant" effects may be politically unwelcome, but editors still prefer the message "nothing works" to the message "everything is uncertain."

By the early 1970s, most social scientists had concluded that if America's goal was to raise student achievement, the policies that most educators favored—smaller classes, better equipment, higher salaries, more extensive teacher training—would not do much good. The bottom line seemed to be that "money doesn't matter." Yet, despite the accumulation of evidence that seemed to point in this direction, neither parents nor educators believed the message. Educators kept asking for more money, legislators kept giving it to them, and the voters mostly went along.

Eventually, educational researchers tired of delivering a message that nobody wanted to hear. Some (including me) turned to other topics. Graduate students in education turned away from quantitative research and began doing qualitative studies, which reduced the risk of finding evidence at odds with

---

[2] See, for example, the assessments in Mosteller and Moynihan (1971).

their prior beliefs about the world. Quantitative educational research did not disappear, but it was marginalized in most schools of education.

The Impact of Meta-analysis

Quantitative educational researchers took a long time to dig themselves out of this hole. The first step was the invention of meta-analysis, which allowed quantitative researchers to pool results from many different studies in a statistically efficient way. When analysts did this, their view of the world changed drastically. Instead of seeing a world in which most studies yielded "statistically insignificant" coefficients, they saw a world in which most small studies yielded coefficients with the expected sign and in which the average coefficient was large enough to be educationally important.

Gene Glass and his colleagues (1982) showed, for example, that when they pooled results from all the available studies of class size, smaller classes were associated with quite large gains in achievement. This was true despite the fact that most of the original studies had reported an "insignificant" relationship between class size and achievement. Once Glass and his colleagues pooled the data, moreover, the relationship was clearly "significant," in the sense that the confidence interval did not include 0. The trouble with the original studies was that they had been too small to provide reliable information about the size of the effect. Meta-analysis of studies assessing the impact of school desegregation told a similar story, at least for elementary school reading achievement (Cook et al. 1984). So did meta-analysis of studies assessing the impact of most other educational policies (Lipsey and Wilson 1993).

These were all classic cases of Type Two error, where earlier analysts had mistakenly rejected the hypothesis that policies had a positive effect. Students had, of course, learned about Type Two errors for generations. But it was not until the advent of meta-analysis that we began to appreciate both the likelihood and the potential costs of such errors.

Indeed, among some quantitative researchers, meta-analysis led to a dramatic paradigm shift. Instead of assuming that "nothing works," they now began to assume that "everything works." But that too was an oversimplification, for several reasons.

Meta-analysis is feasible only when a policy has been studied many times. Thus, the policy has to remain sufficiently popular over a long enough period

to generate numerous studies of its effectiveness. Following Bayesian logic, we would expect policies that have been studied dozens of times to be somewhat more effective than those that have only been studied a few times.

Meta-analysis also highlighted a serious shortcoming of most past educational research. Even when meta-analysts found dozens of studies assessing a particular policy, they seldom found more than a handful of studies that measured the policy's long-term impact. When researchers *had* done long-term follow-ups, moreover, the long-term impact almost always looked smaller than the short-term impact. Among skeptics, therefore, the idea that "nothing works" was gradually replaced by the idea that "everything works in the short run, but benefits usually fade away in the long run."

The idea that achievement gains fade over time may, however, be another artifact of social scientists' statistical conventions. Ever since the invention of IQ tests early in the 20th century, psychometricians have tended to standardize test results. Initially, their goal was to ensure that IQ tests had the same mean and standard deviation at all ages. Standardizing test scores also made it much easier to compare results derived from different tests. But age standardization also obscures a crucial fact about children's cognitive skills, which is that their variance increases with age. If you ask 4-year-olds to do 10 two-digit multiplication problems, their scores are likely to be very similar, because none of the children will be able to do any of the problems. If you ask 14-year-olds to do the same problems, some will get them all right, while others will still get almost all of them wrong. The same logic applies to vocabulary words. The vocabulary of 14-year-olds is more variable than the vocabulary of 4-year-olds.

While this fanning out of academic achievement as children age is well known, meta-analysts usually ignore it. If meta-analysts want to describe the effect of preschool programs at age 4, they will report that those who attended a preschool scored, let us say, 0.30 standard deviations above those who did not attend. If they want to describe the effect five years later, they will report that the gap has shrunk from 0.30 standard deviations to, say, 0.15 standard deviations. They hardly ever ask whether a disparity of 0.15 standard deviations at age 9 is larger or smaller than a disparity of 0.30 standard deviations at age 4.

Meta-analysis has other limitations that curtail its usefulness to policymakers. First, meta-analysts cannot compensate for measurement errors in the studies they pool unless the magnitude of these errors is known. This fact tends to bias estimates of policy impact downward. Second, meta-analysts cannot compensate for the fact that researchers seldom do true experiments, in which randomly selected students are assigned to "treatment" and "control" groups. So-called "natural" experiments, in which students or their parents have a choice about the education they receive, tend to exaggerate the impact of educational policy per se.

Most of the data available to a meta-analyst comes from surveys in which an analyst had identified students affected by some policy of interest, controlled some of the factors correlated with the presence of this policy, and treated any remaining association between the policy and student outcomes as causal. Since analysts can seldom measure all the factors that lead students to have different educational experiences, studies of this kind are likely to suffer from what we once called "omitted variable bias" and now call "selection bias."

Although there is no certain way of eliminating selection bias, better data can often help a lot. The struggle to make quantitative educational research useful to policymakers is critically dependent on these improvements. Fortunately, funders have recognized this. More and more educational surveys now track students over substantial periods of time, measuring both treatments and outcomes on numerous occasions. This strategy provides better estimates of measurement error. It also allows analysts to adjust for the effects of stable unmeasured differences between students. As a result, researchers have been able to generate results that are far more persuasive than those derived a generation ago from the Coleman data or other cross-sectional surveys.

Recent results from analyses of this kind also fit practitioners' expectations better than earlier results did. If this trend continues, researchers may find themselves concluding that conventional wisdom among educators was not as misguided as earlier researchers thought. If researchers can then move on to identifying the most effective strategies for improving achievement, quantitative research may eventually prove quite useful.

The Need for Experiments

Despite all the improvements in survey data and analytic methods, however, there is still one huge problem. Educational researchers used to do

experiments in which students were randomly assigned to treatment and control groups. Today, that is extremely rare. To see why the dearth of experiments poses a problem, consider the debate over ability grouping in elementary schools. This is an intensely controversial issue, especially in racially mixed schools. Yet much of the controversy is about facts that would be easy to ascertain using conventional experimental methods. Suppose we ask the following two questions:

1. Do children who have learned little in the past learn more when they are all assigned to heterogeneous classrooms or when classrooms are segregated on the basis of past academic performance?
2. If students who have learned little in the past are assigned to heterogeneous classrooms, do they learn more when the teacher groups them by skill level or when the teacher treats the entire class as a homogenous group?

If students were randomly assigned to different grouping schemes, it would be fairly simple to see which of these schemes was best for which students. But while experiments of this kind are technically feasible, they are no longer done. When Mosteller, Light, and Sachs (1996) surveyed ability grouping experiments done in elementary schools, they found only one study published after 1980. In politics, data that are more than 20 years old carry little weight.

Nor is it clear that such data *should* carry much weight. Mosteller, Light, and Sachs also found only one experiment that dealt with within-classroom grouping, which is more common and more flexible that between-classroom grouping, at least at the elementary school level. They also found that existing experiments dealt exclusively with mathematics. They found no experiments that reported long-term effects. And while they found little evidence that assigning students to classrooms on the basis of past achievement affected the amount learned when all students were taught the same thing, the question that looms largest in today's debates is what happens when students in "faster" classes cover more material each week, so that they start algebra, geometry, and calculus sooner than they otherwise would.

In theory, it might be possible to answer questions of this kind by comparing school systems that pursue different policies or by following the progress of students who move from one kind of school system to another. In practice, studies of that kind would not be likely to convince skeptics. Experienced school

administrators, board members, and legislators know that when a researcher reports results based on complex multivariate statistics, some other researcher will soon come to the opposite conclusion. Complex statistical analyses require dozens of methodological choices that cannot be made by following a generally accepted rule. In a field as politicized as education, the existence of such choices typically guarantees that nonexperimental data will have many possible interpretations.

Those who analyze experiments are also far more likely to agree about what they show. A legislator or a school board member can follow the logic of the Tennessee class size experiment, understand how the results were evaluated, and see why these results mean what researchers say they mean. Of course, legislators do not understand *exactly* what happened. Nor do they understand all the ways in which the experiment may have been contaminated. Least of all, do they understand the limitations of the findings—that the Tennessee results tell us nothing about the benefits of small classes after third grade, for example. But the structure of the argument is still intuitively obvious to almost everyone.

Given these political advantages, why are randomized experiments so rare? The proximate cause is clear: any randomized experiment disrupts school routines, so educators will participate in one only if they think randomization is absolutely crucial to learning something important. To convince educators that experiments are crucial, researchers must be nearly unanimous in supporting the method. Since 1970, such unanimity has vanished. Surprisingly few educational researchers now see experiments as a good way of going about their business. Indeed, few educational researchers have had any experience with randomized experiments.

One reason most researchers are skeptical about experiments is that they seldom really care whether Policy A is better than Policy B. Most researchers care about *why* Policy A is better than Policy B. It takes a whole series of carefully crafted experiments, each of which rules out one or more alternative explanations, to show *why* something happens. And even after dozens of experiments, there is always the possibility that the experimenter has failed to test a plausible alternative to his or her preferred explanation.

Experiments also have a bad reputation because of the way they were once analyzed. The fact that experiments often seemed to show that all kinds

of policies did not work has inevitably made educators skeptical about the method. We now know that this pattern was partly traceable to the way we analyzed the data and the way we thought about statistical significance. But even today experiments are politically risky, especially when—as is often the case—they are underfunded and therefore too small to detect small effects.

Still, the shortage of experiments is a huge political problem for anyone who thinks that educational policy should be based on evidence. The fact that the Tennessee class size experiment was funded by the state of Tennessee, not the federal government, is indicative of Washington's failure in this area. When the state that brought us the Scopes Trial funds research more useful than that funded by the federal government, something is terribly wrong in Washington.

Recommendations for Future Data Collection

Although I think we ought to be putting far more resources into experiments, that does not mean we should stop doing surveys. It would be as foolish to stop doing surveys and rely exclusively on experiments as it was to stop doing experiments and rely exclusively on surveys, as we did a generation ago. The two methods are complementary.

If we keep doing surveys, as we surely should, we need to think more carefully than we have about what kinds of data we should collect. Brewer and Goldhaber's (1998) list is a good starting point. Some people may be shocked to hear me say this, because theirs is an economist's list, and I am a sociologist. Nonetheless, I like their list.

Collecting Longitudinal Data

First, longitudinal data *is* definitely better than cross-sectional data. Of course, all researchers think longitudinal data is better as long as they do not have to pay the bill. But I am making a stronger argument, namely that longitudinal data yields more knowledge per dollar than cross-sectional data.

But while longitudinal data is better than cross-sectional data, it is not clear that many years of longitudinal data are better than three or four years. If we want to address problems of measurement error and fadeout, we have to follow students for three or four years. But if surveys keep losing 5 percent of their cases every year because students transfer from one school to another and cannot be followed, at least half the sample will be gone after 12 years. In urban systems

where transfers are more common, attrition will be even higher. Furthermore, if a researcher has no findings about older students until her panel of kindergartners has reached college, her funding may dry up in the meantime.

Linking Students to Teachers

Brewer and Goldhaber (1998) are also right about the potential value of linking students to their classroom teachers. If we want to link students to teachers in any meaningful way, however, we also have to collect data in both the fall and the spring, as Meredith Phillips (1998b) argued in her paper for this conference. In addition, we need to ensure that fall and spring testing brackets the school year in a satisfactory way. Testing students in October and April may be convenient, but the school year is more than six months long.

Collecting More Data on Teachers

Brewer and Goldhaber (1998) are also right that collecting more data on teachers is crucial. It is particularly important to gather evidence on the racially charged issue of whether teachers' test scores have a big impact on student achievement, as Ron Ferguson and others have argued (1998). Survey researchers will, of course, have great difficulty testing teachers in today's political environment. That means we need to explore ways of linking our surveys to state records that include teachers' scores on various exams (McLaughlin and Drori 1998).

Domains That Can Be Limited

If we are going to do all this, we also need to identify domains in which we can afford to do less. Surveys planned by committees always have difficulty deciding what to leave out. The committee almost inevitably represents many different interest groups. Indeed, that is usually its main purpose. Such a committee almost inevitably generates a survey instrument that measures many things badly rather than measuring a few things well. As a result, we learn a little about a lot but not much about anything.

That was probably a defensible strategy for the first few national surveys. It is probably not the right strategy for the next century, at least at the secondary level. The broad outlines of what happens in secondary school are now fairly clear. If we are going to collect more data from high school students, we should probably concentrate on one or two topics per survey. At

the elementary level, which has been seriously neglected in the past, we may still need several surveys that tell us a little about a lot before we turn to more focused data collection.

Attitudinal Measures

If we have to cut back on certain kinds of survey items, my suggestions would again be similar to Brewer and Goldhaber's (1998). We have collected many attitudinal measures from students over the past generation. I do not think we have learned much as a result. Students' statements about whether they plan to go to college or have a baby out of wedlock do have some predictive power, which means that people's plans are somewhat stable from one year to the next. But that is hardly news. School-to-school differences in students' *average* responses to attitudinal questions may be a bit more useful, since they may tell us something about school "climate." But aggregating behavioral measures probably tells us far more. What we really want to know is whether policies that seek to change attitudes have a long-term effect. Past surveys have seldom tried to determine what schools were doing to change attitudes.

Sampling Strategies

Another crucial issue is whether to sample more students per classroom. Meredith Phillips' (1998a) analysis of the Prospects data indicates that sampling more students per classroom is probably sensible if the researcher wants to study teacher effects, but not for the study of anything else. At least in Prospects, students in the same classroom turn out to be rather similar, so drawing a large sample drawn from a small number of classrooms yields results with large standard errors.

The Question of Representative Samples

I also agree with the Brewer and Goldhaber (1998) that representative samples have been oversold. It is certainly crucial to have representative samples for *some* purposes, but not for *all* purposes. We should think more carefully about the division of labor in data collection. We need occasional national surveys that gather data on large representative samples, perhaps covering a limited number of domains in depth. But we should be able to do a lot of causal modeling with state and local data that are not perfectly representative of any well-defined universe.

The data that state testing programs now collect is of limited interest to most researchers, because states do not gather much background information from students or much program information from schools. But if we could match state data on school achievement to survey data on a subset of students in each school, the result might be ideal for researchers interested in causal modeling. Imagine having the kind of data that Texas collects on individual students linked to the kind of data that NELS collects. Since Texas collects data over a student's entire school career, at least as long as the student stays in Texas, researchers could use such data to answer all kinds of questions they cannot answer now.

Unresolved Problems

I want to close by posing some questions that I think we still need to address if we are to make educational research more relevant to educational policy.

What Can We Learn from Differences between States?

For nearly 100 years, progressive policymakers have claimed that decentralized decision-making gave America an unusual opportunity to learn about the effects of different policies. The states, we were constantly told, were the "laboratories of democracy." In education, however, variations in state policy have taught us surprisingly little. We did learn one thing from the American states' diverse educational experiences in the 20th century, which was that de jure racial segregation had terrible consequences. But we eliminated de jure segregation a generation ago.

Since 1970, the 50 American states have continued to pursue 50 different sets of policies. As far as I can tell, we have learned nothing from this diversity. In part, this is because we have not collected good information about what policies states were really pursuing. In part it is because we have not collected good information about how educational outcomes differed from state to state. Primarily, however, our failure to learn from states' experiences reflects the fact that learning from such experiences would require detailed data linking year-to-year changes in educational policy to changes in subsequent outcomes. We simply do not have such information.

The papers that David Grissmer (1998) and Steve Raudenbush (1998) presented this morning suggest that NAEP may now provide state-level out-

come measures that can offer useful policy guidance. Grissmer certainly drew some interesting lessons from North Carolina and Texas. But Raudenbush found that a handful of variables explain almost all state-to-state differences in student achievement. If that is true, these laboratories of democracy may be rather like a child's chemistry set that includes only a handful of different chemicals. This does not imply that educational policy is unimportant. But it does imply that states may be so internally heterogeneous in the policies their schools pursue that state means on the policy variables that matter are relatively similar. This issue requires further exploration.

How Should We Define and Measure School Achievement?

Educators try to teach specific skills and information. If we want to make research useful to educators, we have to measure the skills and information that educators try to teach. We also have to measure skills and information on scales that allow us to say how much students have learned between one period and the next. That requires two major changes in the way educational researchers go about their business.

First, we probably have to stop equating the quality of a test with its reliability. The way to get high reliability is to choose items that are highly correlated with one another. But in a world with diverse schools, teachers, and curriculums, tests with high inter-item correlations almost always end up measuring general ability, not the specific skills and information that educators in particular places have tried to impart.

The other far-reaching change we will have to make is to stop standardizing tests to predetermined means and variances. I do not think we will ever make much progress in measuring learning if we keep thinking about achievement in exclusively relative terms, as we mostly have for the past 100 years.

To see how misleading relative rankings can be, consider the Tennessee class size experiment. Tennessee assigned children to either large or small classes from kindergarten through the end of third grade. The children assigned to smaller classes did better at the end of kindergarten. They preserved, but did not widen, their advantage over the next three years. Many people have been puzzled by the fact that smaller classes did not seem to yield any further benefit after the first year of the experiment. Skeptics like Eric Hanushek (1999) have interpreted this finding as evidence that only the first year in a small class yields measurable benefits. But, as Jeremy Finn pointed out in the discussion

today, this picture reflects the fact that those who analyzed the Tennessee data reported treatment effects at different ages as a percentage of the standard deviation at that age.[3] When one looks at unstandardized treatment effects, they increase as time goes on.

How Can Educational Research Serve Non-policy Goals?

Up to this point I have discussed educational research as if its only legitimate purpose was to improve educational policy by helping policymakers decide whether small classes are worth the extra cost, whether to group students on the basis of their past achievement, and so on. But educational research has another function as well. It tells us something about how we are doing as a people and as a nation. It helps us compare our performance both to our ideals and to the performance of other democratic societies. As a result, it plays a significant role in our judgments about whether we live in a just or an unjust society, whether opportunity is more equal in our society than in other societies, and whether things were really better a generation ago, when most of us were growing up. These are not policy questions in any ordinary sense. But how we answer these questions has an important impact on how we think about ourselves and what policies we favor or oppose.

If American students learn less math than Japanese students, for example, it is not at all clear what policy implications this fact should have. But we still want to know the fact. Similarly, when the black-white test score gap falls dramatically, as it apparently did during the 1980s, this brute fact does not tell us anything about *why* the gap fell or what policies might further reduce it in the future.[4] But even when facts of this kind have no clear policy implications, they tell us something about what is happening to our country that we should—and do—care about.

If I am right in claiming that one major goal of educational research is to tell ordinary citizens how well their country is doing in various domains, we need to report such information in a form that most citizens can understand. I have already argued that comprehensibility is a major argument for preferring randomized experiments to multivariate statistics. The need for comprehensi-

---

[3] Comments from Jeremy Finn during the discussion period of the seminar, November 9, 1998.

[4] See Grismer, Flanagan, and Williamson (1998).

bility should also play a central role in the way we report data on states and school districts. One example must suffice. Researchers often report mean scores for states or school districts that have been adjusted statistically to eliminate the effects of demographic differences between states or districts. This kind of adjustment poses many statistical problems. But its most serious defect may be that it is very hard to explain. Thus, wherever possible, it is probably better to present separate means for different kinds of students: those whose parents have had different amounts of schooling, for example, or those from different racial and ethnic groups. Readers can then see for themselves how each group fares in different states.

Can We Predict the Future?

If anyone had asked me in the late 1960s what educational researchers would learn over the next 15 years, I would have predicted far more progress than actually occurred. It is stunning how little progress we really made between the late 1960s and the early 1980s. But if anyone had repeated the question in the early 1980s, I would again have been wrong. It would never have occurred to me that after 15 years of spinning our wheels, we were about to make progress at an unprecedented rate. Yet that is what happened. The papers presented at this seminar are, I think, much better than anyone would have predicted 15 years ago.

We should all reflect on this history and try to use it to identify the preconditions for intellectual progress. Why, after an apparently promising start in the 1960s, did we accomplish so little in the next 15 years? Why did things go so much better over the next 15 years? I have my own hunches. Meta-analysis forced us to rethink our approach to significance tests. NCES began collecting better data. Econometricians pushed us to adopt more rigorous standards of proof. But these are speculations. A careful intellectual history of quantitative educational research remains a task for the future.

References

Brewer, D. J., and Goldhaber, D. D. (1998). Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 169–188). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Cook, T., Armor, D., Crain, R., Miller, N., Stephan, W., Walberg, H., and Wortman, P. (1984). *School Desegregation and Black Achievement*. U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Institute of Education.

Ferguson, R. F., and Brown, J. (1998). Certification Test Scores, Teacher Quality, and Student Achievement. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 133–156). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Glass, G., Cahen, L., Smith, M. L., and Filby, N. (1982). *School Class Size*. Thousand Oaks, CA: Sage Publications.

Grissmer, D. W., and Flanagan, A. (1999). Moving Educational Research toward Scientific Consensus. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 43–90). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Grissmer, D.W., Flanagan, A., and Williamson, S. (1998). Why Did the Black-White Test Score Gap Narrow in the 1970s and 1980s? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 182–228). Washington, DC: Brookings Institution Press.

Hanushek, E. A. (1999). The Evidence on Class Size. In S. E. Mayer and P. E. Peterson (Eds.) *Earning and Learning: How Schools Matter*. Washington, DC: Brookings Institution Press.

Jencks, C. (1971). The Coleman Report and the Conventional Wisdom. In F. Mosteller and D. P. Moynihan (Eds.), *On Equality of Educational Opportunity* (pp. 69–115). New York: Random House.

Lipsey, M., and Wilson, D. (1993). The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-analysis. *American Psychologist 48*(12): 1181–1209.

McLaughlin, D., and Drori, G. (1999). School-level Correlates of Reading and Mathematics Achievement in Public Schools. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 189–236). U.S. Department of Education. National Center for Education Statistics. Washington, DC:

Mosteller, F., Light, R., and Sachs, J. (1996). Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size. *Harvard Educational Review 66*: 797–842.

Mosteller, F., and Moynihan, D. P. (Eds.) (1971). *On Equality of Educational Opportunity*. New York: Random House.

Phillips, M. (1998). *Do African-American and Latino Children Learn More in Predominantly White Schools?* Unpublished manuscript, School of Public Policy and Social Research, University of California, Los Angeles.

Phillips, M. (1999). Understanding Ethnic Differences in Academic Achievement: Empirical Lessons from National Data. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 103–132). (U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Raudenbush, S. W. (1999). Synthesizing Results from the NAEP Trial State Assessment. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 3–42). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Improving Research and Data Collection on Student Achievement[1]

Brenda J. Turnbull, Policy Study Associates; David W. Grissmer, RAND; and J. Michael Ross, National Center for Education Statistics

This seminar brought together a diverse group of people linked by their motivation to understand and improve student achievement. Participants in the seminar included researchers whose work has focused on analyzing student achievement, policymakers instrumental in designing policies to improve student performance, and government officials who design and manage the collection of major data sets used by researchers and policymakers. The seminar sought to lower the inevitable communication barriers existing within this community: between researchers and policymakers, between those designing and collecting data and those who use it; and among researchers from different disciplines analyzing different data sets with different models and estimation techniques. From the divergent perspectives, we sought to identify directions for future research and data collections, and perhaps a common conceptual framework encompassing research and data collection on achievement.

Here we summarize the recommendations made by participants for more sophisticated data collection strategies, new directions for future research, and collaborative forums for communication of research results. These recommendations fall into two broad categories. The first category includes a variety of smaller incremental changes focusing on improving nonexperimental results, while the second category includes more radical departures from current directions in federal statistical agencies. Finally, we focus on a different topic of discussion at the seminar—communication of results—and provide some concluding remarks on future directions.

[1] The authors are grateful to Martin Orland, John Ralph, Joseph Conaty, and Daniel Kasprzyk for an early discussion of the implications that could emerge from this seminar. However, the opinions expressed herein are those of the authors; and no endorsement by the National Center for Education Statistics, the Office of Educational Research and Improvement, or RAND should be inferred.

Improving Nonexperimental Research and Data Collection

Improving longitudinal surveys was the focus of many recommendations. The National Center for Education Statistics (NCES) and the Department of Education fund several key longitudinal surveys, such as the National Education Longitudinal Survey (NELS:88) and the Early Childhood Longitudinal Studies (ECLS). Since achievement tests are included in their designs, during the last 20 years these surveys have supported a significant amount of research on achievement. Seminar participants could—and did—disagree over the advantages of specific analytic procedures, but none questioned the essential value of having and expanding the collection of longitudinal data. The recommendations for improving existing longitudinal studies centered on the inevitable competing priorities for specific topics and survey questions, the frequency of data collection, more stratified sampling plans, and the importance of prekindergarten baseline measures, as well as a more structured process for the design of all surveys. Dominic Brewer and Dan Goldhaber presented a list of specific recommendations that also received support from Christopher Jencks; indeed, most participants contributed recommendations in this area.

One key issue in survey design was highlighted by an exchange from the floor. Jeremy Finn suggested that a survey should focus on only six to twelve well-specified constructs. Jencks countered that a government agency cannot make such a draconian selection, because it is answerable to many constituencies. Jencks did suggest, though, that NCES could take a retrospective look at the actual uses made of particular variables and particular items, saying that this kind of analysis would support the selective deletion of less productive items from repeated surveys. This inductive procedure would be quite different from the deductive one proposed by Finn, where a clear and bounded conceptual framework would drive the construction of survey items. These remarks reflect the reality that a large-scale survey qualifies both as research and as a political undertaking, because the data ultimately are used for many purposes. However, scientific constructs can improve the items devoted to research issues and at the same time perhaps constrain the usually high demand for items devoted to nonresearch issues.

An allied issue is the discontinuation of past survey items in order to include new items. From time to time during the day, a participant would suggest ways of trimming the length of surveys by deleting specific items.

Generally, another participant would swiftly object, arguing for the theoretical or practical significance of those items. An example was the suggestion by Brewer and Goldhaber that "school climate" measures could be substantially reduced, which was followed by a rejoinder from Valerie Lee that these measures are integral to an assessment of school quality. Such disagreements over items partly reflect the still unresolved disciplinary preferences for the importance of certain variables and modes of analysis. These disagreements suggest that design teams should include researchers and scholars from different disciplines insofar as they can engage in productive dialogue over these issues from their different perspectives.

Suggestions were also made concerning the frequency and timing of collecting longitudinal data. Collecting data each year rather than every two years when following a cohort through school was mentioned, but the increased costs or trade-offs with other survey design parameters would certainly need to be considered. Meredith Phillips made a compelling case for shifting the annual testing of students to testing in both fall and spring. Although burdensome in implementation, this change would permit analysis and comparison of students' learning trajectories during the regular school year and the summer. The importance of this issue was demonstrated in Phillips' analysis, which confirms that the test score gap between black and white students widened during the summer time period. Her recommendation was echoed by Smith, Jencks, and, from the floor, Adam Gamoran.

Recommendations were also made with respect to sampling strategies relevant to classroom-within-school effects. Brewer and Goldhaber and Lee supported nested samples with more students per classroom. Lee suggested that more students per class could be efficiently traded off for fewer schools in longitudinal samples. In addition, Brewer and Goldhaber, along with Ron Ferguson and Jordana Brown, suggested sampling more teachers per school. Realistically, these changes would increase costs and burden, but their endorsement by researchers does suggest that these trade-offs should at least be given serious consideration in the design of future samples.

Several presenters emphasized the desirability—as well as the challenges—of finding out "what teachers know." Stephen Raudenbush mentioned the importance of measuring teachers' subject matter preparation and content knowledge in their assigned teaching fields. For Ferguson and Brown, teachers' scores on tests such as the ACT are also important data, although Ferguson

said that NCES should "probably not" try to collect such data itself, but should instead play a facilitating and convening role with private sector organizations that already collect such data. Brewer and Goldhaber suggested such possible measures as administering a written assessment to teachers or asking principals to assess teacher quality.

Finally, David Grissmer and Ann Flanagan suggested that longitudinal surveys need to begin at school entry in order to capture the variables needed for accurate estimation of both short- and long-term effects in student learning. They suggest that production function methodologies that utilize only a previous test score as a proxy for earlier schooling variables are challenged by the results of the Tennessee experiment. They observe that these results suggest a multiyear effect from class size reductions, and these multiyear effects cannot be controlled for by a single-year previous test score. Therefore, survey resources should be shifted to earlier years, and the value of surveys started at later grades would be diminished.

Improving NAEP Data

Data from the National Assessment of Educational Progress (NAEP) provide the only achievement data with representative samples of U.S. students. NAEP has been administered to samples of students at ages 9, 13, and 17 since 1969 in several subjects. While the trend results from these data have been often used and cited in many studies, research explaining the trends is more recent. The NAEP data remain the most reliable information for assessing the changing gap between minority and white students, and also for facilitating inquiries into the question of whether educational and social policies directed toward minority and low-income families and students have raised student achievement test scores. The expansion of NAEP, since 1990, to state samples also provides the only comparative achievement scores across states with representative samples within each state. Several recommendations for improving the NAEP data emerge from the research presented in the seminar.

Grissmer and Flanagan observe that the most serious weakness of NAEP can be found in the student-reported family characteristics. They suggest implementing a simple parental survey as one means of collecting better data. However, they also suggest testing empirically whether student-reported surrogate measures such as "books in the home" can provide adequate substitutes

for parent-collected data. They also suggest using state-level Census data to supplement NAEP family data.

However, any significant addition to the NAEP data collection such as a parent survey may have significant opportunity costs. Sylvia Johnson and Gary Phillips mentioned the current press to speed up and simplify NAEP, pointing out that this works against the more complex data collection that would permit us, in Johnson's words, to "better understand the whys and hows of improving student achievement."

Johnson, as well as Grissmer and Flanagan, made the more radical suggestion of changing the sampling plan from school samples to school district samples. Several advantages might accrue from a district-based sample, although the more complex sampling would increase costs. A district-based sample would allow comparing and explaining achievement differences across major urban school systems relative to smaller, more homogeneous suburban districts. Urban school systems encompass a significant part of the nation's education problems, but we currently have no adequate comparative measures of performance across these urban systems. Johnson observed that the change to school district sampling could help spur improvement in learning because it would heighten public scrutiny of districts' results. However, if costs were kept constant, such sampling would probably mean fewer students per school, allowing more variance in measurement of school characteristics.

Recommendations for Changing Directions

Pointing to the near universal credibility enjoyed by Tennessee's experiment with reduced class size, several presenters called for more experiments (randomized field trials) to evaluate education programs. Deputy Secretary Marshall Smith characterized the policy impact of the Tennessee experiment as "instantaneous" and "incredibly powerful in Congress." Compared with conventional surveys, he said, such experimental trials provide more robust and probably more valid estimates of program effects. Also emphasizing the power of experiments to communicate, Jencks credited experiments with producing results that are easy to understand, saying "The structure of the argument is intuitively obvious." Grissmer and Flanagan argued that research consensus is more likely to emerge when a well-designed, -implemented, and -analyzed experiment has taken place and when the analysis can show little sensitivity to

the inevitable deviations from ideal design specifications. Although nonexperimental studies can be filtered through the scrutiny of meta-analysis and expert panels, they observed that this process often does not lead to consensus.

Recognizing the high costs and the limits of experiments, several speakers also addressed strategies for accumulating evidence over time through a series of inter-related investigations. Smith cited the studies related to the acquisition of reading skills supported by the National Institute of Child Health and Human Development (NICHD), which he described as "not a single experiment, but a series of interlocking studies testing hypotheses about reading." These studies enabled NICHD to "tell a very coherent story about reading, with immense power" in policy circles, and thus they have paved the way for the appropriations needed to further experimental studies. Jencks, in answer to a question from the floor, endorsed the idea of embedding small experiments within more conventional survey designs.

Deputy Secretary of Education Smith proposed several topics which would benefit from experimental investigations, each representing an area in which there is a theoretical and an evidentiary base upon which to design randomized trials, as well as strong policy interest:

- ***After-school and summer programs.*** The rationale for offering extended learning time is clear, but little or no good evidence is yet available on the effects of well-designed programs that provide students with a safe environment and adult tutoring beyond the typical school day and year.
- ***Training in parenting skills.*** Such training could be combined with adult education in an experimental trial.
- ***Education-focused preschool programs***. Evidence now indicates that children benefit from acquiring particular skills and concepts before they enter school, such as knowing the alphabet, knowing that one reads from left to right, knowing the concepts of *before* and *after*, *up* and *down*, and the like. The effects of such instruction could be studied experimentally.
- ***Research-based interventions in school reform.*** David Cohen of the University of Michigan is launching a major study of schoolwide reforms. His methods involve detailed survey and observational methods rather than an experimental test of the policy that encourages schools to implement well-specified programs under the guidance of outside

experts. Despite the current enthusiasm for schoolwide reforms, they have not been well studied in either education or business.

What could not be explored in much depth in the discussions, however, was the place that experiments might occupy in an overall portfolio of support for research and how decisions would be made concerning such a portfolio of experiments. Jencks noted that the Tennessee experiment was launched by the state of Tennessee without federal help, and the state of Wisconsin has also initiated a quasi-experiment with student-teacher ratios. Both NCES and the Office of Educational Research and Improvement (OERI) could face uphill struggles, for somewhat different reasons, if they were to try to sponsor experiments. For a statistical agency to take the lead in initiating an experiment would be unusual and might jeopardize the stability of its role and mission, at least according to some observers. A well-funded research agency can include experiments in its portfolio, but OERI currently lacks the discretionary resources to launch such costly investigations. This dilemma suggests the need for an innovative federal-state partnership in the organization and implementation of an experimental agenda.

In the last 10 years, state testing programs have become the major source of achievement data in the nation. Virtually every state is now committed to more frequent testing of its students statewide in a variety of subjects across a variety of grades. The largest representative sample collected in national achievement data is approximately 25,000 students, and the largest state samples collected by NCES (NAEP Trial State Assessment) usually test around 2,500 students per state. Samples for state-administered tests include nearly all students with the exception of certain IEP (Individual Education Program) and LEP (Limited English Proficient) students who are excluded, so sample sizes can be over one million students in several states. National tests are typically given every four years, while state testing often occurs on an annual basis within certain grade ranges. Moreover, in some states, individual student scores can be tracked across grades and linked to specific teachers, thereby allowing even richer longitudinal and contextual analysis.

This explosion of achievement testing in many states suggests a new direction for federal data collection effort; namely, using state achievement data as a platform for research and experimentation. The McLaughlin and Drori study provides an example of linking this state assessment data at the school level with federal data such as the Schools and Staffing Survey (SASS). State

achievement data can also be supplemented with state teacher data, more detailed information on resources, and facility information. The Ferguson and Brown analysis illustrates the power of linking these data with teacher characteristics and other data available in Texas. Their paper drew from state databases that permit analysis of teacher test performance in relation to student performance at the district level in Texas. Ferguson recommended more such work, but observed that NCES would not have to administer tests to teachers in order to obtain data. Instead, Ferguson suggested that the federal government play a more active role of leadership in convening and coordinating research efforts with organizations such as Educational Testing Service (ETS) and perhaps with state agencies through the exchange of data with appropriate protections for individual confidentiality.

Yet the potential for a richer "universe" as sources of data may lie in evaluating planned interventions and experimentation. Many states already have many ongoing education reforms that could be better evaluated with these data sets. However, a federal leadership role would encourage both random assignment of schools or districts by states in the initial phase of program implementation and then the funding of high quality program evaluations conducted by national experts.

Both controlled field trials and longitudinal studies are usually seen as intrinsically complex and costly endeavors. However, such research can range from small to large scale. Departing from the general endorsement of loading more complexity onto one large multipurpose study, Robert Hauser suggested that an alternative would be the more frequent initiation of smaller longitudinal studies, such as the Wisconsin Longitudinal Study. These would differ from the "larger, one-time-only or once-per-decade surveys" that have been customary in education. He elaborated:

> We ought to be initiating cohort surveys close to birth every year—or every other year—as a means of improving our "who, what, when?" understanding. Such surveys should be stratified by ethnic origin, differentially sampled. And they should provide opportunity for experimentation with alternative test (and questionnaire) content and observational designs, as well as for core content stable enough to permit aggregation of findings across cohorts to yield greater statistical power.

Such small-scale but in-depth studies can lead to better model specifications in larger studies, even if they are not totally representative of the entire population. Hauser argues that it is hard to successfully model achievement in large samples, until we understand more precisely the development of individual students in different contexts. Such smaller scale studies, if focused on testing the assumptions usually made in larger scale research studies, could probably contribute significantly to the development of more coherent research findings.

Small-scale experiments are also possible, given that achievement data are usually already collected across many different states. Different types of interventions are occurring in schools, and early development models may be easily turned toward at least a quasi-experimentation orientation in the early stages of implementation. Universities in each state could become centers for initiating intervention research and experimentation that builds on current assessment data. In discussing research on existing databases, Ferguson and Brown, based upon their experience in Texas and Alabama, emphasized the value of working with researchers who are based in a state and have a long-term career interest in working with that state's database. Again, the federal role would be to convene researchers around a shared knowledge-building agenda, rather than simply to supply data.

Communicating Research Results

The seminar's recurring emphasis on the communication of new findings often challenged conventional assumptions and served as a reminder that statistical and explanatory presentations are an integral part of more sophisticated methods of data collection and advances in data analysis techniques. Further, decisions about the specific form of a presentation can make a difference in the persuasive power and ultimate value of data and research for policymakers.

Smith described the power of the Tennessee experiment to influence policymakers—a result Jencks attributed to the transparency of experimental results. Jencks noted the quite different implications that are often communicated by displaying aggregate achievement data versus the same data when it is disaggregated into different demographic groups. For instance, the display of aggregate NAEP trends has often been used to imply that no gains have occurred in achievement over the last 25 years, while the display of minority trends shows significant gains in achievement, at least up to 1988.

Communicating nonexperimental results is more difficult due to the nontransparency of the analysis. However, Raudenbush's paper provides a compelling example of how a display of nonexperimental results illustrates the varying factors related to achievement differences between states. The graphical display of achievement differences across states illustrates the differential resources and opportunities across states. In both previous and the current work, Raudenbush has also advocated "value-added" models of achievement when evaluation of schools or teachers is involved. Separating the impact of family and social capital from specific schooling effects is a part of being able to effectively partition the components of achievement differences and effectively communicate results.

Some other participants also addressed the topic of communication. Grissmer advocated more emphasis on support for professional consensus activities through conferences, National Research Council panels, and the preparation of edited books focused on specific topics.[2] Support for consensus panels, which are common in health research, was expressed, the intent being to provide a more formal basis to study, form consensus, and communicate important research conclusions.

Concluding Remarks

Reviewing the seminar papers in their totality, we are encouraged by the large and increasing amount of achievement data being collected—some longitudinal, some cross-sectional—and we are quite optimistic that some persistent and perplexing research questions are now empirically answerable. Besides the data collected by NAEP and the combination of main and State NAEP, it is now possible to determine the relative variations between states and between schools within states. In addition, the research is now beginning to examine in more detail the resource differences associated with these outcome measures. In addition, some states have collected longitudinal databases at the student level for all schools and districts, therefore permitting researchers to measure grade-specific changes in student performance over their years in school. Many of these new databases also allow the statistical examination of multilevel factors (such as school- and district-level factors) and social context factors, where the performance of certain types of students may be quite different depending

[2] Examples of recent books include the following: Ladd 1996; Burtless 1996; Jencks and Phillips 1998; Ladd, Chalk, and Hansen 1999; and Ladd, Chalk, and Hansen 2000.

upon the social and racial composition of the school. On the horizon, data will be available for a new longitudinal cohort of students beginning in kindergarten and extending through fifth grade, which are coupled with parent, teacher, and administrator surveys; these new data sets should fill in many gaps in our knowledge concerning instructional practices and the relative importance of family and out-of-school social factors. Finally, we are optimistic that, in the near future, new governance arrangements will expand opportunities for state and federal researchers to link and analyze new databases, thereby generating more refined statistical estimates of these factors.

The accurate and valid measurement of student achievement and performance will increase in importance over time. Not only is "testing" becoming more "high stakes" in some states, but research on these issues is also becoming "high stakes." For better or worse, student achievement is increasingly being used to measure the effectiveness of schools in some states, as well as the overall effectiveness of the resources committed to education change. The federal government, state legislatures, and district policymakers increasingly utilize the findings from this research to guide and justify their policies; and research findings are commonly utilized in equity and adequacy lawsuits being pursued in many states. Likewise, the "quality of education" ranks near the top of voters' concerns in national, state, and local elections; and, unfortunately, candidates for public office are tempted to utilize achievement test scores as evidence for the success or failure of new reform policies. Significant increases in educational spending are a common part of the agenda of both political parties, and these increases are now often linked to accountability for producing higher achievement.

Although the findings from education research are being cited and utilized more frequently by policymakers, a wide range of inconsistent, and at times contradictory, research results are often put forth. Researchers have many different explanations for why findings are inconsistent. One explanation attributes the inconsistency to the actual difference in the effects across different contexts. This explanation trusts the modeling process to accurately estimate a "real" effect, but expects the results to differ due to the different contexts. For instance, two measurements of the effects of per pupil expenditure might differ because the money may be allocated more effectively in some cases, and the inconsistency is interpreted as reflecting a public school system that lacks incentives to utilize money well.

A second explanation attributes the inconsistent results to flaws in the modeling process, and not to "real" differences in effects. It assumes that the inconsistent findings reflect imprecise modeling methods that do not fully reflect the complexity of the process being measured. Previous models are rarely comparable, because different data sets are used with varying quality of data or different model specifications or different estimation methods. In this case, widely varied findings are interpreted as inconclusive, and conclusions are not drawn from these models about the efficiency or effectiveness of school systems.

It is difficult to determine how much of the inconsistency is due to imprecision in modeling versus real contextual variation in the actual effect when we rely primarily on nonexperimental measurements, but several directions for research can be helpful. The first is to use similar model specifications and estimation across a variety of large data sets to eliminate certain sources of variance—different specifications and estimation methods.[3] A second direction is to utilize data sets with the most complete sets of variables to explore the sensitivity of variable coefficients to less complete variable sets.[4] (For instance, many databases contain few family background variables, often just race-ethnicity and a measure of income obtained at one point in time, such as qualifying for free and reduced price lunch. Data sets containing a much richer set of family resource variables can be used to explore which schooling variables are affected by less complete sets of family variables and how much sensitivity exists.) A third direction is for literature reviews to address directly the question of why many findings differ across studies rather than simply note that findings are widely dispersed.[5] A fourth direction is to use different estimation techniques across the same data sets with similar variables to determine how sensitive these measures are to different estimation methods. A fifth direction is to empirically test directly some of the major assumptions used in

---

[3] One example of this is Hedges and Nowell (1998). They fit models with similar family variables across a variety of data sets that have achievement and family variables.

[4] An example of this is Goldhaber and Brewer (1997).

[5] A recent example of this is Krueger (1999). Krueger takes the findings from each study used in Hanushek, Rivkin, and Taylor (1996) and tests whether the conclusions are sensitive to the criteria used for measurement inclusion and level of aggregation. Other examples include Greenwald, Hedges, and Laine (1996) and Hanushek (1996), where the sensitivity of results to output measures, specification, time of publication, school level, and level of aggregation are tested. Grissmer et al. (forthcoming) also try to explain the pattern of different results of previous measurements by level of aggregation.

nonexperimental analysis.[6] There are examples in the literature of each of these five types of analysis, but there is not a coherent effort to improve our nonexperimental analysis based upon these studies.

Focusing on why nonexperimental measurements produce different findings and discovering the underlying hypothesis that explains the pattern of variance can be helpful, but it is doubtful whether research consensus can emerge from this process alone. There are simply too many different possibilities that might explain why nonexperimental measurements differ and current data sets have significant limitations in their ability to support such analysis. However, two complementary directions—experimentation and micro-process theory building—may eventually provide decisive evidence that can explain the variance in nonexperimental analysis and lead to a convergence of evidence from which scientific consensus can emerge. Well-designed, -implemented, and -analyzed experimental data that show little sensitivity to the inevitable deviations from ideal design can serve two purposes. First, more reliable measurements can serve as benchmarks for evaluating results from nonexperimental models. Second, experimental data can serve to calibrate nonexperimental models by identifying specifications and estimations that can predict experimental results. Thus, experimental data can contribute to our understanding of why inconsistent findings exist in nonexperimental measurements and why convergence between experimental and nonexperimental evidence may eventually emerge.

Still, finding consistency across experimental and nonexperimental evidence leaves out one important element from which scientific consensus develops. The power of theories is that they can successfully explain past empirical results and predict future empirical results and ultimately provide the authority from which scientific consensus emerges. Theories that successfully predict how teachers and students will change behavior in smaller classrooms, and how that changed behavior leads to higher achievement, and why there are different behavior changes in classes with high and low SES begin to generate the authority for consensus. The ultimate test, however, is for the theory to generate new constructs and operational definitions and subsequently to predict their effects on achievement outcomes.

---

[6] An example of this is Heckmaan, Layne-Farrar, and Todd (1996).

Theory building does not receive much emphasis in education research, partly because it is difficult terrain to negotiate. Theory building is necessarily multidisciplinary because the building blocks come from psychology, child development, cognitive science, economics, and sociology. Further, it involves human behavior as well as developmental processes. However, theory building can begin with much simpler goals, such as predicting how achievement changes when resources or teachers change by linking changes in classroom and home behavior when resources change. Successful theories are not only an ultimate goal of good scientific research, but also can be very effective communication tools since they tell "stories" of why resource changes affect achievement.[7]

Theory development requires a coordinated research agenda and more comprehensive data collections. The Tennessee class size experiment, with its associated collection of classroom data, was a good beginning for the 1980s. But this experiment could have collected better data aimed at explaining why achievement was sustained after students left the smaller classes in later years in school. What behavior or developmental pattern changed to sustain these results through eighth grade, and why did those students in small classes for only one to two years not sustain gains as compared to those in small classes for three to four years? More detailed classroom and home observations, more information concerning peer and family relationships, and eventually brain processing patterns in early and later grades for students in small and large classes can help explain such effects, and provide the basis for theory building.

Reflecting on the presentations, discussions during the seminar, and the revised papers in this book, we are hopeful that the deliberations have enabled federal researchers and other policymakers to take stock of innovative research, and especially of achievement research, as areas of empirical inquiry. This process of rethinking basic assumptions should facilitate new understandings of where education research has been in the past and, more importantly, where new challenging research opportunities may be presenting themselves in the future. Thus, our efforts to ensure richer quality of information from our data sources, improved methods of empirical inquiry, and more informative theory building should be enhanced by an occasion when different groups come together to exchange ideas and present written summaries of their findings. It is,

---

[7] See *Educational Evaluation and Policy Analysis 20*(2) (1999, summer) on class size for an example of linking research together to support theory building.

for example, gratifying to realize the abundance of data now available, compared to the skimpy information prior to the publication of "the Coleman report" in 1966. Further, it is encouraging to recognize the expanding applications to education policy that are the results of utilizing federal and state data collections. For the diverse research community encompassed in the seminar, intellectual challenges remain: fine-tuning existing data collection strategies, exploring linkages and connections between federal and state data sources, improvements in specifications within statistical models, and building relevant and useful theories of education processes. Thoughtful consideration of selected recommendations contained in this report should lead to a more coherent and productive research agenda for federal statistical agencies and state research organizations; to promising partnerships between these researchers and independent entities such as ETS; and eventually to an expansion of knowledge that facilitates and promotes student learning in schools.

References

Burtless, G. (1996). *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington, DC: The Brookings Institution Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., York, R. L. (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

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

Greenwald, R., Hedges, L. V., and Laine, R. D. (1996, fall). The Effect of School Resources on Student Achievement. *Review of Educational Research 66*(3): 361–396

Grissmer, D. W., Flanagan, A., Kawata, J., and Williamson, S. (forthcoming). *Improving Student Achievement: State Policies That Make a Difference.* Santa Monica, CA: RAND.

Grissmer, D. W. (1999, summer). Assessing the Evidence on Class Size: Policy Implications and Future Research Agenda. *Educational Evaluation and Policy Analysis 20*(2).

Hanushek, E. A., Rivkin, S. G., and Taylor, L. L. (1996, fall). Aggregation and the Estimated Effects of School Resources. *The Review of Economics and Statistics*, pp. 611–627.

Heckman, J., Layne-Farrar, A., and Todd, P. (1996). Does Measured School Quality Really Matter? In H. F. Ladd (Ed.) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (pp. 192–289). Washington, DC: The Brookings Institution Press.

Hedges, L. V., and Nowell, A. (1998). Group Differences in Mental Test Scores: Mean Differences, Variability, and Talent. In C. Jencks and M. Phillips (Eds.) *The Black-White Test Score Gap.* Washington, DC: The Brookings Institution Press.

Jencks, C. and Phillips, M. (1998). *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

Krueger, A. B. (1999). *An Economist Looks at Class Size Research.* Working Paper. Princeton, NJ: Princeton University.

Ladd, H. F. (Ed.). (1996). *Holding Schools Accountable.* Washington, D.C.: The Brookings Institution Press.

Ladd, H. F., Chalk, R., and Hansen, J. S. (1999). *Equity and Adequacy in Educational Finance: Issues and Perspectives.* Washington, DC: National Academy Press.

Ladd, H. F., Chalk, R., and Hansen, J. S. (2000). *Making Money Matter.* Washington, DC: National Academy Press.

Section V.
Appendix

Seminar Attendees *(as of November 1998)*

Nancy Allen
Educational Testing Service
Rosedale Road, Mailstop 02T
Princeton, NJ 08541
E-mail: nallen@ets.org

Judith Anderson
U.S. Department of Education
National Institute on Student Achievement, Curriculum, and Assessment
555 New Jersey Ave., NW
Room 517
Washington, DC 20208
E-mail: judith_anderson@ed.gov

Pat Anderson
District of Columbia Public Schools
Department of Educational Accountability
825 N. Capitol Street, NE
Room 8053B
Washington, DC 20002

Mark Berends
RAND
1333 H Street, NW
Washington, DC 20005–4707
E-mail: mark_berends@rand.org

Sue Betka
U.S. Department of Education
Office of Educational Research and Improvement
555 New Jersey Ave., NW
Room 600G
Washington, DC 20208–5649
E-mail: sue_betka@ed.gov

Eve Bither
U.S. Department of Education
National Education Research, Policy &
Priorities Board
555 New Jersey Ave., NW
Room 101
Washington, DC 20208

Rolf Blank
Council of Chief State School Officers
One Massachusetts Ave., NW
Suite 700
Washington, DC 20001–1431
E-mail: rolfb@ccsso.org

George Bohrnstedt
American Institutes for Research
P.O. Box 1113
Palo Alto, CA 94302–1113
E-mail: gbohrnstedt@air-ca.org

Geoffrey D. Borman
The Johns Hopkins University
3003 North Charles Street
Suite 200
Baltimore, MD 21218
E-mail: gborman@csos.jhu.edu

Anne Bouie
U.S. Department of Education
Office of Educational Research and Improvement
555 New Jersey Ave., NW
Suite 611c
Washington, DC 20208
E-mail: anne_bouie@ed.gov

Mary Lyn Bourque
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002
E-mail:
mary_lyn_bourque@ed.gov

Dominic Brewer
RAND
1333 H Street, NW
Washington, DC 20005–4707
E-mail:
dominic_brewer@rand.org

Jordana Brown
John F. Kennedy School of Government
Malcolm Wiener Center for Social Policy
79 John F. Kennedy Street
Cambridge, MA 02138
E-mail: jordana_brown/
fs.ksg@ksg.harvard.edu

Janis Brown
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 318
Washington, DC 20208
E-mail: janis_brown@ed.gov

Peggy G. Carr
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 308
Washington, DC 20208
E-mail: peggy_carr@ed.gov

Chris Chapman
U.S. Department of Education
National Center for Education Statistics
Room 422K
555 New Jersey Ave., NW
Washington, DC 20208
E-mail: chris_chapman@ed.gov

Lisa Chavez
University of California at Berkeley
16A Forest Street, #41
Cambridge, MA 02138
E-mail:
llchavez@uclink4.berkeley.edu

Susan Choy
MPR Associates
2150 Shattuck Avenue
Suite 800
Berkeley, CA 94704
E-mail: schoy@mprinc.com

Arthur L. Coleman
U.S. Department of Education
Office for Civil Rights
330 C Street, SW
Room 5012D
Washington, DC 20202

Joseph Conaty
U.S. Department of Education
National Institute on Student Achievement,
Curriculum, and Assessment
555 New Jersey Ave., NW
Room 510H
Washington, DC 20208
E-mail: joseph_conaty@ed.gov

Lynn Cornett
Southern Regional Education Board
592 10th Street, NW
Atlanta, GA 30318–5790
E-mail: lynn.cornett@sreb.org

James Crouse
University of Delaware
Educational Studies
213H Willard Hall
Newark, DE 19716
E-mail: jcrouse@udel.edu

Nancy Doorey
Delaware State Board of Education
4601 Beechwold Avenue
Wilmington, DE 19803
E-mail: ndoorey@den.k12.de.us

Kimberley Edelin
Frederick D. Patterson Research Institute
UNCF
8260 Willow Oaks Corporate Drive
Fairfax, VA 22031
E-mail: edelink@fdpri.patterson-uncf.org

Ronald Ferguson
John F. Kennedy School of Government
Malcolm Wiener Center for Social Policy
79 John F. Kennedy Street
Cambridge, MA 02138
E-mail:
ron_ferguson@harvard.edu

Michael Feuer
National Research Council
Board on Testing and Assessment
2101 Constitution Avenue, NW
Washington, DC 20418
E-mail: mfeuer@nas.edu

Jeremy Finn
State University of New York
Baldy Hill
Room 408C
Amherst, NY 14260
E-mail: jfinn@acsu.buffalo.edu

Pascal D. Forgione, Jr.
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Suite 500
Washington, DC 20208
E-mail: pascal_forgione@ed.gov

Norm Fruchter
Institute for Education and Social
Policy
New York University
726 Broadway-Fifth Floor
New York, NY 10003

Edward Fuentes
U.S. Department of Education
National Institute on the
Education of
At-Risk Students
555 New Jersey Ave., NW
Room 610E
Washington, DC 20208
E-mail: edward_fuentes@ed.gov

Adam Gamoran
University of Wisconsin
Department of Sociology
1180 Observatory Dr.
Madison, WI 53706
E-mail: gamoran@ssc.wisc.edu

Margaret Goertz
University of Pennsylvania
Consortium for Policy Research
in Education
3440 Market Street
Suite 560
Philadelphia, PA 19104–3325
E-mail: pegg@wfs.gse.upenn.edu

Daniel Goldhaber
The Urban Institute
2100 M Street, NW
Washington, DC 20037

Edmund Gordon
3 Cooper Morris Drive
Pomona, NY 10970

Steven Gorman
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 404G
Washington, DC 20208

David Grissmer
RAND
1333 H Street, NW
Washington, DC 20005–4507
E-mail: david_grissmer@rand.org

Eric Grodsky
University of Wisconsin
Sociology Department
2210 Sommers Avenue
Madison, WI 53704
E-mail: egrodsky@ssc.wisc.edu

Kerry Gruber
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 422B
Washington, DC 20208
E-mail: kerry_gruber@ed.gov

Maureen Hallinan
University of Notre Dame
400 Decio
Notre Dame, IN 46556
E-mail: hallinan.i@nd.edu

Jane Hannaway
The Urban Institute
2100 M Street, NW
Washington, DC 20037
E-mail: jhannawa@ui.urban.org

Robert Hauser
University of Wisconsin
Center for Demography & Ecology
1180 Observatory Drive
Madison, WI 53706
E-mail: hauser@ssc.wisc.edu

Larry V. Hedges
University of Chicago
5835 Kimbark
Chicago, IL 60637
E-mail:
hedge@cicero.spc.uchicago.edu

Lori Diane Hill
University of Chicago
OSC/NORC
1155 East 60th Street
Room 385
Chicago, IL 60637
E-mail: lori@cicero.spc.uchicago.edu

Emily Howard
U.S. Department of Education
Office of the Deputy Secretary
600 Independence Ave., SW
Washington, DC 20202

Christopher Jencks
John F. Kennedy School of Government
Malcolm Wiener Center for Social Policy
79 John F. Kennedy Street
Cambridge, MA 02138
E-mail: jencks@wjh.harvard.edu

Jackie Jenkins
U.S. Department of Education
National Institute on Student Achievement,
Curriculum, and Assessment
555 New Jersey Ave., NW
Room 611B
Washington, DC 20208

Christopher Johnson
University of Michigan
125 West Hoover Ave.
Apt. 1B
Ann Arbor, MI 48103
E-mail: cjque@umich.edu

Sylvia Johnson
Howard University
2900 Van Ness Street, NW
116 Holy Cross Hall
Washington, DC 20008
E-mail: sjohnson@howard.edu

Will Jordan
The Johns Hopkins University
3505 N. Charles Street
Baltimore, MD 21218
E-mail:
wjjordan@jhunix.hcf.jhu.edu

Dan Kasprzyk
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 422H
Washington, DC 20208
E-mail: daniel_kasprzyk@ed.gov

Andrew Kolstad
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 406B
Washington, DC 20208
E-mail: andrew_kolstad@ed.gov

Daniel Koretz
Center for Testing
Boston College
Campion Hall
Room 3223
Chestnut Hill, MA 02167
E-mail: daniel_koretz@rand.org

Karol Krotki
American Institutes for Research
Education Statistics Services Institute
1000 Thomas Jefferson St., NW
Suite 400
Washington, DC 20007
E-mail: kkrotki@air-dc.org

Valerie Lee
University of Michigan
School of Education
610 East University Avenue
Room 4220
Ann Arbor, MI 48109–1259
E-mail: velee@umich.edu

Sharon Lewis
Council of Great City Schools
1301 Pennsylvania Avenue, NW
Suite 702
Washington, DC 20004
E-mail: slewis@cgcs.org

Serge Madhere
Howard University
Psychology Department
Washington, DC 20059
E-mail: smadhere@howard.edu

Jo Anne Manswell
Howard University
CRESPAR
2900 Van Ness Street, NW
Holy Cross Hall, Room 427
Washington, DC 20008
E-mail:
jmanswell@crespar.law.howard.edu

John Mazzeo
Educational Testing Service
Rosedale Road
Mail Stop 30-E
Princeton, NJ 08541
E-mail: jmazzeo@ets.org

C. Kent McGuire
U.S. Department of Education
Office of Educational Research
and Improvement
555 New Jersey Ave., NW
Room 600
Washington, DC 20208

Don McLaughlin
American Institutes for Research
P.O. Box 1113
Palo Alto. CA 94302–1113
E-mail: dmclaughlin@air-ca.org

Marilyn McMillen
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 406A
Washington, DC 20208
E-mail:
marilyn_mcmillen@ed.gov

James McPartland
The Johns Hopkins University
CSOS
3003 N. Charles Street
Suite 200
Baltimore, MD 21218

Joseph McTighe
Council for American Private
Education
18016 Mateny Road
Room 140
Germantown, MD 20874
E-mail: cape@connectinc.com

Stephen Morgan
Harvard University
Department of Sociology
William James Hall, 5th Floor
Cambridge, MA 02138
E-mail:
smorgan@wjh.harvard.edu

Ann L. Mullen
U.S. Department of Education
National Center for Education Statistics Fellow
555 New Jersey Ave., NW
Room 406
Washington, DC 20208
E-mail: ann_mullen@ed.gov

John Mullens
Policy Studies Associates
1718 Connecticut Avenue, NW
Suite 400
Washington, DC 20009
E-mail:
jmullens@policystudies.com

Jennifer O'Day
University of Wisconsin-Madison
Department of Educational Policy Studies
1000 Bascom Mall
225 Ed. Bldg.
Madison. WI 53706
E-mail:
joday@mail.soe.madison.wisc.edu

Michael Olneck
University of Wisconsin-Madison
School of Education
1000 Bascom Mall
Madison, WI 53706
E-mail:
olneck@mail.soe.madison.wisc.edu

Martin Orland
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 402D
Washington, DC 20208
E-mail: martin_orland@ed.gov

Eugene H. Owen
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 402E
Washington, DC 20208
E-mail: eugene_owen@ed.gov

Jeffrey Owings
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 417A
Washington, DC 20208
E-mail: jeffrey_owings@ed.gov

Audrey Pendleton
U.S. Department of Education
Planning and Evaluation Service
600 Independence Ave., SW
Washington, DC 20208
E-mail:
audrey_pendleton@ed.gov

Robert Petrin
University of Chicago
OSC/NORC
1155 E. 60th Street
Chicago, IL 60637
E-mail:
robp@cicero.spc.uchicago.edu

Gary Phillips
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 408C
Washington, DC 20208
E-mail: gary_phillips@ed.gov

Meredith Phillips
UCLA
School of Public Policy and
Social Research
Public Policy Building
Room 3250
Los Angeles, CA 90098–1656
E-mail: phillips@sppsr.ucla.edu

Paul Planchon
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 413B
Washington, DC 20208
E-mail: paul_planchon@ed.gov

Valena White Plisko
U.S. Department of Education
Planning and Evaluation Service
600 Independence Ave., SW
Room 4143
Washington, DC 20208
E-mail: valena_plisko@ed.gov

John Ralph
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 402L
Washington, DC 20208
E-mail: john_ralph@ed.gov

Stephen Raudenbush
University of Michigan
School of Education
610 East University Avenue
Ann Arbor, MI 48109–1259
E-mail: rauden@umich.edu

Karen Ross
University of Michigan
1072 Island Drive
Apt. 105
Ann Arbor, MI 48105
E-mail: keross@umich.ed

Michael Ross
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 412
Washington, DC 20208
E-mail: michael_ross@ed.gov

Richard Rothstein
Economic Policy Institute
P.O. Box 301
South Welfleet, MA 02663

Laura Salganik
American Institutes for Research
Education Statistics Services Institute
1000 Thomas Jefferson Street, NW
Suite 400
Washington, DC 20007
E-mail: lsalganik@air-dc.org

Alex Sedlacek
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 404B
Washington, DC 20208
E-mail: alex_sedlacek@ed.gov

Ramsay Selden
American Institutes for Research
Education Statistics Services Institute
1000 Thomas Jefferson Street, NW
Suite 400
Washington, DC 20007
E-mail: rselden@air-dc.org

Sharif Shakrani
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002
E-mail: sharif_shakrani@ed.gov

Becky Smerdon
American Institutes for Research
Pelavin Research Center
1000 Thomas Jefferson Street, NW
Suite 400
Washington, DC 20007
E-mail: bsmerdon@air-dc.org

Marshall S. Smith
U.S. Department of Education
Office of the Deputy Secretary
600 Independence Avenue, SW
Room 6236
Washington, DC 20202

Holly Spurlock
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Avenue, NW
Room 408A
Washington, DC 20208
E-mail: holly_spurlock@ed.gov

Christopher B. Swanson
University of Chicago
OSC/NORC
1155 East 60th Street
Chicago, IL 60637
E-mail: cb-swanson@uchicago.edu

Ricky Takai
U.S. Department of Education
Planning and Evaluation Service
600 Independence Avenue, SW
Room 4111
Washington, DC 20202
E-mail: ricky_takai@ed.gov

Corrine Taylor
University of Wisconsin-Madison
5202 Tolman Terrace
Madison, WI 53711
E-mail: ctaylor@ssc.wisc.edu

Bill Trent
National Academy of Sciences
University of Illinois at Urbana-Champaign
Swanlund Administration Building
601 E. John Street
Champaign, IL 61820
E-mail: w_trent@uiuc.edu

Suzanne Triplett
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Avenue, NW
Room 408D
Washington, DC 20208
E-mail: suzanne_triplett@ed.gov

Roy Truby
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002

Clyde Tucker
U.S. Department of Labor
Bureau of Labor Statistics
2 Massachusetts Avenue, NE
Washington, DC 20212
E-mail: tucker_c@bls.gov

Brenda Turnbull
Policy Studies Associates
1718 Connecticut Avenue, NW
Suite 400
Washington, DC 20009
E-mail:
brenda@policystudies.com

Rob Warren
University of Washington
Department of Sociology
202 Savery Hall
Box 353340
Seattle, WA 98195–3340
E-mail:
jrwarren@u.washington.edu

Walter G. West
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Avenue, NW
Room 417B
Washington, DC 20208
E-mail: walter_west@ed.gov

Paul Williams
Educational Testing Service
Rosedale Road
Mail Stop 30-E
Princeton, NJ 08541
E-mail: pwilliams@ets.org

John Wirt
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Avenue, NW
Room 402F
Washington, DC 20208
E-mail: jwirt@inet.ed.gov

Alexander Wohl
U.S. Department of Education
Office of the Secretary
600 Independence Avenue, SW
Room 6101
Washington, DC 20202

Shi-Chang Wu
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Avenue, NW
Room 415
Washington, DC 20208