

**METHODOLOGICAL CONSIDERATIONS AND
INTERPRETATION OF THE RESULTS OF
DIFFERENT MODELS**

The state NAEP scores are the first achievement scores with representative sampling across states that allow comparison of performance across states. Since states are the leading initiators of educational reform, these scores may be the principle source for evaluating the effectiveness and cost-effectiveness of reform initiatives. These scores are also important because across-state variation accounts for two-thirds of the total variance in per-pupil expenditures among school districts. Because the reasons for variations in expenditures across states may differ from the reasons for variations in expenditures within a state, measurements of resource effects across states are of interest.

However, a number of analytical issues must be addressed in using these data, and some caution is needed in interpreting the results for policy evaluation. One issue is whether to use the individual-level or the aggregate state-level data for evaluation. Both methods should eventually be used, and their results should be compared. Here, we present the aggregate results and suggest several reasons that aggregate-level results may eventually prove less biased than individual-level analysis with NAEP data. The reasons are as follows:

- Higher-quality family variables are available at the state level.
- Higher-quality schooling variables related to major resource expenditures are available at the state level, since they can include estimates from entry into school.

- Evidence from previous measurements of the effects of educational resources shows more-consistent and robust results at the state level of aggregation, and these results are in closer agreement with experimental data than are individual-level results.
- There are no a priori reasons to prefer analysis at the state or individual level, since results should agree if models are linear and well specified.
- There are no a priori reasons to suggest that the net bias introduced by various types of misspecification will be higher for state-level analysis than for individual-level analysis.
- For some forms of misspecification that we know are present (selection, measurement errors, and missing variables at the individual level), there are plausible reasons to suggest that individual-level results will be more biased than state-level results.

Three potential problems with aggregate data are a small sample size, a limited range of variation in variables, and heightened sensitivity to outliers compared to individual-level data. Aggregate models—provided a sufficient range of variation is present in variables—can screen and detect any variables that have large effects but often cannot discriminate variables with weaker effects. Aggregate analysis is less likely to produce statistically significant results and less able to detect low-level effects for resource variables. So, these models should be used to measure effects for major resource variables only.

Insignificant effects may also need a different interpretation for aggregate models. Statistical insignificance may be due to limited sample size and range of variation rather than absence of an actual effect. However, statistically significant results for individual and aggregate results can be interpreted in a similar way.

We specifically addressed whether our results are sensitive to certain specification issues by estimating random- and fixed-effect models that each depend on different, but plausible, sets of statistical assumptions. Since family variables explain much of the variance across state scores, we developed three different sets of family variables using different sources of family data to test the sensitivity of the results to different family variables. Finally, we tested for the influence of outliers on the results.

We generally found consistency in estimates across different model specifications and family variables and little sensitivity to outliers. The fixed-effect models provide coefficients that generally show similar or somewhat smaller effects compared to the random-effects models. Not unexpectedly, the fixed-effect results are also less robust, given our limited time series and the reduced degrees of freedom in these models. The random-effect models provide the most consistent effects and show the closest agreement with the Tennessee results.

Among the six models estimated (random and fixed effect for the three sets of family variables), we found very similar results for the state trends. One set of estimates began to show somewhat different results from the other five estimates for the value-added and policy models. In this model, the family coefficients began to show perverse signs. This model—the fixed-effect estimates with Census-NAEP family variables—also has the fewest degrees of freedom. We have reported these results in the appendixes but have not included them in the main results.

As a test of model validity, we made predictions of the pattern of results from the Tennessee class-size experiment using the 4th-grade scores in our sample. We used only random-effect models because of the reduced sample size of the 4th-grade sample. The three random-effect models predicted results that are very consistent with the size of the Tennessee effects and also the pattern of increasing effects for minority and lower-SES students.

RESULTS

The pattern of achievement across states from 1990 to 1996 suggests that family variables explain most of the variance across scores in states. So, raw scores are of little value as indicators of the effectiveness of the K–12 educational system in states. Our results also show the presence of significant state-specific effects on achievement that make scores differ across states for students with similar family background. Some of these differences in scores can be explained by specific characteristics of the educational systems in the states. These results suggest that the level of per-pupil expenditures and how they are allocated and targeted can make significant differences in student achievement. The results also suggest that there is a wide

variation in the cost-effectiveness of expenditures, depending on the grade level, the programs targeted, and the SES level of students.

Finally, the results suggest that significant gains in math scores occurred from 1990 to 1996, but there is wide variation in gains across states. Our resource variables cannot explain much of the overall gains or the pattern of gains by state.

More specifically, the results show the following:

- Public elementary students across participating states had statistically significant gains in mathematics of about 1 percentile point a year between 1990 and 1996.
- The rate of math improvement varied significantly across states with a few states making gains of about 2 percentile points a year, while other states had little or no gain.
- These gains cannot be explained by changes in levels or allocation of the major resource variables (per-pupil expenditure, teacher salaries, pupil-teacher ratio, teacher resources, and levels of public prekindergarten participation). Differences in systemic reform initiatives across states are the leading candidates for explaining these gains in scores.
- Students from similar family backgrounds had statistically significant score differences across states that can be as large as 11 to 12 percentile points. States with both higher and lower scores for students from similar families are in all regions of the country.
- The characteristics of state educational systems can account for some of the score differences for students from similar families. Other things being equal, higher per-pupil expenditures, lower pupil-teacher ratio in lower grades, higher reported adequacy of teacher-reported resources, higher levels of participation in public prekindergarten, and lower teacher turnover all show positive, statistically significant effects on achievement. Other things being equal, higher teacher salaries, higher teacher educational levels, and increased experience over the past three years do not show significant effects on achievement. However, these variables explain one-half or less of the nonfamily differences in achievement across states.
- The cost-effectiveness of expenditures varied widely, depending on how the expenditures are directed, the SES level of the state,

the current allocation of expenditures, and the grades targeted. The most efficient uses of educational expenditures among those evaluated here were providing all K–8 teachers more-adequate resources for teaching, expanding public prekindergarten in lower-SES states, and targeting reductions in pupil-teacher ratios in lower grades in lower-SES states to well below the national average and in medium-SES states to the national average. Estimates of the cost-effectiveness of teacher aides from Tennessee experimental data show that they are far less cost-effective than class-size reductions.

Evidence for the Effects of Reform

The effects of the wave of reform initiatives that began in the later 1980s would be expected to begin showing up in achievement scores in the 1990 to 1996 period. But it will take much longer for the full effects to be reflected. These reform initiatives are of two types: Some are related to changes in the level and allocation of resources, while others are more structural or pedagogical. Since the pace and structure of reform vary considerably by state, such efforts would be expected to show uneven results across states. However, only gains that are unrelated to traditional resource variables would provide evidence for the effect of structural reform.

The analysis provides strong evidence that math scores from 1990 to 1996—controlling for student demographic changes and NAEP participation rates—increased in most states for public students by statistically significant amounts. Eighth-grade math scores increased much more than 4th-grade scores. The average annual gain across states is about 1 percentile point a year. A few states gained about 2 percentile points a year, while some states had little or no gain.

Three sources of evidence point to structural reform as the cause of these gains, but much more research is needed into the pattern of specific reforms across all states before a compelling linkage can be made. First, these gains cannot be explained by major changes in resources. Second, the rate of gains varied widely by state, as did the structure and intensity of reform efforts. Third, a case study of two states with the largest gains—Texas and North Carolina—suggested that a series of similar reforms in both states linked to aligned stan-

dards, assessments, and accountability was the most plausible cause of the gains.

Scores for Students from Similar Family Backgrounds

Since achievement is strongly linked to family characteristics, only score differences for students with similar family backgrounds indicate the presence of state-specific effects that might be linked to the quality of educational systems and other factors. Our results indicate the strong presence of differences in scores across states for students from similar families. The analysis was able to distinguish three groups of states: those whose students with similar family backgrounds are significantly above and below the median state, and a broad middle group. Texas and California—two states with fairly similar family characteristics—are ranked the highest and lowest, respectively, and have score differences for students with similar family backgrounds of about 12 percentile points. The variables in our model explain about two-thirds of the difference in these scores. The major contributions to the higher Texas scores are lower pupil-teacher ratio, a much larger percentage of children in public prekindergarten, and teachers who report having more of the resources necessary to teach.

The Effects and Cost-Effectiveness of Resources

States have widely different levels of per-pupil expenditures, which are spent in significantly different ways across states. The result is wide variance in pupil-teacher ratios, teacher characteristics, and salaries; different levels of teacher-reported adequacy of resources; and different support for public prekindergarten programs. Such large differences would be expected to show up in achievement scores across states if these variables actually affect achievement in significant ways. States also have wide variation in the SES characteristics of their students, so it is possible to measure whether these resource variables have different effects across SES levels.

The results imply that resources in public education must be allocated to specific programs and grade levels and toward specific students to be most effective and cost-effective. The cost-effectiveness of resource expenditures can change by more than a factor of 25,

depending on which programs and grade levels are funded and which types of students are targeted. The analysis suggests that providing all K–8 teachers additional resources, expanding prekindergarten in low-SES states, reducing pupil-teacher ratios in lower grades in lower-SES states to well below the national average, and reducing pupil-teacher ratio in medium-SES states to the national average are most efficient. This analysis suggests that significant gains in achievement for students in lower-SES states can be achieved through modest increases in resources, if allocated to specific programs. Conservative estimates show predicted score gains of 12 to 15 percentile points from additional targeted expenditures of less than \$1,000 dollars a pupil in the states with the lowest SES.

INTERPRETATIONS

Improving American Education

The most widely accepted explanation of the pattern of previous measurements on the effects of educational resources has been that public education shows great inconsistency in the effectiveness with which it uses additional resources. This explanation does not rule out the occasional effective use of resources but generally proposes that public schools have used additional resources ineffectively and inefficiently. One explanation for this ineffectiveness is that sufficient incentives do not exist in the public school systems to drive effective use of resources.

The major evidence that supports this explanation comes from previous empirical measurements of the effect of a wide variety of resources made at the individual, classroom, school, and district levels that show highly inconsistent results. While the overall results suggest a net positive effect for resources, the wide inconsistency is interpreted as indicating that public schools do not reliably and effectively convert resources into better outcomes.

This explanation implies that providing more resources for public education is not the answer to school improvement without fundamental reforms that can change the organizational climate and incentives in education. An underlying thesis is that the public school system is too bureaucratic to reform itself and that it is necessary to create alternatives outside the current system or increased

choice within the system to produce an environment of greater competition. Policies advocated with this approach include vouchers, school choice, charter schools, and contracting out of schools.

Recent research has suggested three major problems with this explanation. First, there were significant score gains nationwide for minority and disadvantaged public school students in the 1970s and 1980s. This period was characterized by social and educational policies focused on these groups and by disproportionate allocation of relatively modest levels of additional educational resources to programs that primarily benefited these groups. Second, the results of the Tennessee class-size experiment showed large and significant effects from class-size reductions in public schools, and these results appear robust to the inevitable flaws in experimental design and execution. Third, previous measurements at the state level of aggregation, unlike lower levels of aggregation, showed consistent statistically significant positive effects from additional resources. The usual explanation given is that state-level effects are biased upward with respect to the more-accurate measurements made at lower levels of aggregation, but no credible source of bias has been suggested.

We suggest a competing explanation for the pattern of results in the previous literature that is consistent with the results from the Tennessee experiment, the pattern of national score gains and expenditures from 1970 through 1996, and the new results in this report. This explanation suggests that measurements at the state level may provide the most-accurate results among previous measurements and that less-aggregate measurements may be biased downward. The newly available state NAEP scores provided a test of whether using much-higher-quality achievement data at the state level could also provide effects consistent with past state measurements. We found positive and statistically significant effects from higher levels of resources across states and found that certain resource uses can be highly effective and efficient in raising achievement.

These results also suggest that additional resources are most effective and efficient when spent in states with higher proportions of minority and disadvantaged students. Thus, these results suggest that the modest additional resources spent in the 1970s and 1980s might account for significant national minority and disadvantaged score gains in this period. In particular, national pupil-teacher ratios declined significantly, and evidence from our model and from the

Tennessee experiment would suggest that such reductions are consistent with explaining part or most of the gains in the 1970s and 1980s. Our state NAEP model also provides pupil-teacher effects consistent with the size and pattern of larger effects for minority and lower-SES students found in the Tennessee experiment.

We also suggest that specific forms of bias are known to exist in educational data that could plausibly bias previous measurements made at the individual, classroom, school, and district levels downward but that these introduce much less bias at the state level of aggregation. One form of such bias is clearly indicated by the Tennessee experimental data: missing variables on schooling conditions from school entry. These variables are missing in almost all previous measurements and would probably create a larger bias in a less-aggregate analysis, likely in a downward direction.

The Tennessee data also suggest that production-function models that contain a pretest measure—generally thought to be the highest-quality specifications and often used at less-aggregate levels of analysis—are untenable and can lead to significant bias, likely in a downward direction. Selection effects—widely acknowledged to exist in education—also have the potential to produce more bias at less-aggregate levels.

This explanation does suggest a different approach to improving public education. The public-education system can and has used some additional resources effectively, particularly when directed to minority and disadvantaged students. Our results suggest that such resources need to be effectively targeted to specific programs, matched to particular types of students, and toward early grades. Targeting resources toward states with lower-SES students appears to be the most efficient. The current disparity in per-pupil spending across states represents a source of major inefficiency in educational spending.

Our results also show that significant gains are occurring in math scores across most states, with sizable gains in some states. The source of these gains cannot be traced to resource changes, and the most likely explanation would suggest that ongoing structural reform **within** public education might be responsible. This reform suggests that well-designed standards linked to assessments and some forms of accountability may change the incentives and productivity within

public schools and even introduce competition among public schools. Thus, these results certainly challenge the traditional view of public education as “unreformable.” Public education may be a unique type of public institution in which competition and accountability work because of the large number of separate units whose output can be measured.

There are reasons to believe that improvements in achievement could be expected to continue. The full effect of structural reform initiatives is not reflected in current achievement, and the identification of successful initiatives will likely result in diffusion across states. Better allocation of future resources can also raise achievement. Finally, new data, improving methods of nonexperimental analysis, and new experimentation could also be expected to contribute to future gains.

Interpreting the Effects of Teacher Salary and Teacher Working Conditions

The variables in our analysis that are most efficient seem to involve improving the classroom teaching environment or “working conditions” for teachers. Smaller pupil-teacher ratios and higher levels of satisfaction with resources for teaching appear to make teachers more productive. Prekindergarten participation may improve the classroom environment by providing better-prepared students for teachers. However, our equations imply that, other things being equal, states having higher average salaries do not have higher achievement.

This analysis would suggest that salary increases might come at the expense of providing teachers the working conditions that make them more productive. The analysis suggests that, if investments were made to improve teacher working conditions in the ways recommended, the **current** teachers in our schools would produce significant gains in achievement scores. The Tennessee experiment also supports the conclusion that changes in the conditions facing teachers in the classroom result in higher achievement. The efforts to increase the quality of teachers in the long run are important, but this analysis would suggest that significant productivity gains can be obtained with the current teaching force if their working conditions

are improved. It further suggests that teachers by and large respond to better conditions and know what to do to raise achievement.

The low cost-effectiveness of direct investment in salaries can have at least four interpretations. The first explanation assumes the measurements are accurate and attempts to explain the ineffectiveness of increases in teacher salary. The second explanation is that the teacher salary coefficient is biased downward because of its correlation with family variables. The third explanation posits that measurements of interstate salary differences may show different effects from measurements of intrastate salary differences. The fourth interpretation attributes the weak salary effect to the excess supply of teachers in the 1980s and early 1990s.

The ineffectiveness of teacher compensation could result from the inefficient structure of the current teacher compensation system and the inability to target salary increases to higher-quality teachers effectively (Grissmer and Kirby, 1997; Hanushek, 1994; Ballou and Podgursky, 1995, 1997). Unlike class size, which can be targeted to early grades and lower-SES students, salaries are, by and large, raised for all teachers. If the system could distinguish and provide higher compensation for higher-quality teachers and those who are more effective with lower-scoring students, for whom there is more leverage for raising scores, one would expect a dollar of compensation to be more effective. However, the differential pay by school districts in the current salary system is insufficient to prevent higher-quality teachers from teaching in districts with higher-SES students.

A second problem is that salary increases are usually given for more education and experience. Teacher educational level is a weak and inconsistent predictor of achievement. For universities and colleges, providing teachers with master's degrees produces significant income but seems to have little effect on improving teachers' abilities to raise achievement. Teachers themselves are motivated to spend significant time and money on pursuing such degrees largely because of the structure of the current compensation system. It is arguably one of the least-efficient expenditures in education.¹

¹More master's degrees in education are awarded annually than in any other subject, constituting one in four master's degrees awarded in the nation, with more than 100,000 awarded annually. Assuming the cost of tuition, transportation costs, and the

Teacher experience shows somewhat stronger and more consistent results, but other teacher characteristics generally show more-consistent relationships with achievement. Verbal ability, test scores of teachers, and degrees in subjects taught are three such characteristics, and others may exist. So, part of the ineffectiveness of the current compensation structure is that pay is not directed toward rewarding characteristics that are related to producing higher achievement.

The second explanation is that the coefficient of teacher salary is biased downward because of its correlation with social capital. An overlooked source of social capital can be teachers, who are usually seen as part of the schooling effect. If teachers disproportionately choose to teach in or are more often hired to teach in schools whose students have family characteristics similar to their own, the teachers must be considered as part of social capital. It is almost certainly the case that teachers from backgrounds with more family resources are more likely to teach students with more family resources and vice versa. This is partly due to the fact that teachers usually teach in the same state in which their own schooling occurred, often returning to the same city and county of their own schooling. Hiring practices that often attempt to match the characteristics of students and teachers probably reinforce this trend.

Thus, correlation probably exists between teacher characteristics and student characteristics. The highest correlation between family and school characteristics is between teacher salary and family characteristics (approximately 0.60). If the characteristic of teachers that determines their effectiveness partially has its origin in family capital (i.e., verbal ability), part of the effects of higher-quality teachers may appear in the social-capital effect.

However, two effects are possible when teachers and students are matched. If effective teaching has a component linked to intrinsic characteristics correlated with the teacher's family background (i.e., verbal ability), matching teacher and student characteristics will have net positive effects on achievement of students in families with more resources. However, a second effect can arise if teachers can

opportunity costs of time, a master's degree conservatively costs \$20,000. Annual national expenditures by teachers or subsidized by school districts would be approximately \$2 billion.

more effectively teach students from backgrounds similar to their own: the mentoring effect.

Given the current situation, in which teachers are more likely to be matched to student backgrounds, both of these effects would be positive for students with high family resource backgrounds. For students from lower family resource backgrounds, one effect would be positive and one negative, and whether the net effect is positive or negative would depend on the relative strength of the two. Regardless of which effect dominates, the net effects are likely to be captured in the social-capital effect unless specific variables are introduced measuring the intrinsic abilities of teachers (teacher test scores) and the characteristics of the match between student and teacher.

A third explanation is that the effects of interstate salary differentials may be different from intrastate differentials. Teachers tend to teach in their home states and may be sensitive to salary differentials across districts within a state but are less sensitive to salary differentials across states. Part of the reason may be that women constitute over two-thirds of the teaching workforce and do not have the same job mobility as men in seeking higher-paying jobs. Thus, intrastate differences in salary may affect the distribution of quality teachers much more than the interstate salary differentials. In this case, an intrastate analysis may show salary to be more effective in increasing achievement.

Last, more sensitivity to salary would be expected when teacher labor markets are tight. The period of this analysis was characterized by a teacher surplus across most teacher categories. However, the teacher labor market will become much tighter from 2000 to 2010 for several highly predictable reasons. An aging teacher force will have increasing retirement rates, and attrition rates will also stay high if outside job opportunities remain plentiful. Reductions in class size will also likely increase demand for new teachers.

The supply of new teachers depends partially on the labor market for college graduates, which has been strong in recent years. Beginning teacher salaries are not competitive with most alternative job opportunities. Thus, it may be difficult to expand supply without significant salary growth.

A teacher shortage disproportionately affects schools in lower-SES districts, where the leverage is greatest for boosting scores but also where the risk is greatest for achievement declines. So, the tightening teacher labor market might be expected to heighten the sensitivity of achievement to salary levels—especially for lower-SES states and localities.

RESEARCH IMPLICATIONS

It would not be surprising if some educational resources had not been used effectively in education because policymakers and educators have had little help from the research and development (R&D) community in identifying what is effective and efficient. Successful R&D is the engine that drives productivity improvement in every sector of our economy. Until educational R&D can play the role that R&D does in virtually every other sector of our economy, continual educational improvement cannot be taken for granted.

Experimentation

More experimentation in education seems critical. However, in the long run, confidence in nonexperimental results is needed for policy guidance, since only a limited number of experiments are possible, and contextual effects will likely be important influences in education. Thus, the generalizability of experimental data may always be limited, and we will have to depend on improved nonexperimental analysis. Therefore, experimentation should be directed not only toward variables that have major resource implications but also toward hypotheses that can significantly improve our specifications with nonexperimental models.

Obtaining accurate estimates of resource and policy effects is only the first step needed for policy guidance. The second is to estimate the costs accurately and to compare the cost-effectiveness across resource uses and policies. Cost analysis needs to be built into experimentation, and nonexperimental analysis needs to be more focused on cost-effectiveness analysis.

Improving Nonexperimental Analysis

Besides experimentation focused on testing assumptions in nonexperimental analysis, there are several research directions to improve

the reliability of nonexperimental analysis. Use of individual-level longitudinal data that begin at school entry can sort out many of the specification problems that may exist in previous analyses. There are two new sources of such longitudinal data that will have school, teacher, and family characteristics and achievement data. First, there are newly emerging longitudinal state databases that link student achievement across years. Such data have very large samples, and linkages are possible with teacher data and school characteristics. These data should help sort out many of the potential specification issues involving dependence of later achievement on previous years' class sizes and thresholds and on interactions with teacher characteristics. However, certain forms of bias may still be a problem with individual-level data, even if it is longitudinal from kindergarten.

It will also likely be possible to determine class-size effects for various combinations of large and small classes in early and later grades and the importance of small classes in later grades. The subject of differential bias across levels of aggregation can also be partially addressed with these data through direct testing.

The second source will be the Early Childhood Longitudinal Study funded by the U.S. Department of Education, which will collect very detailed data on children, their families, and their schools. The data will be much richer in variables but will have much smaller sample sizes.

A second approach to improving the reliability of nonexperimental analysis is to use empirical analysis to test and better understand the assumptions upon which such analysis depends. Why do students in large and small classes have different characteristics? How important are parent and teacher selection processes in determining class size? Do more-senior teachers choose smaller classes? Are assumptions more valid in some kinds of schools? Are class sizes in rural areas mainly randomly determined, with more selection occurring in cities? There are many empirical approaches to addressing these kinds of questions that would give us a better idea whether assumptions made in specifications are reasonable.

Finally, it now appears that specifying models will require more knowledge about classroom behavior and children's cognitive development. Neither the classroom nor the child can be treated as a

black box. There is a great deal of research on patterns of physical, emotional, and social development in children from birth, covering such areas as differences across children, delays in development, and dependence on previous mastery. Studies involving long-term developmental outcomes—especially for children at risk—identify resiliency factors that enable development to occur even in highly risky situations. Much can be learned from this literature to help prevent the use of poor modeling assumptions.

Building Theories

Experimentation and improved nonexperimental analysis alone will not build scientific consensus. Theories need to be developed that link classroom behavior and its effect on student development with resource variables. Theories that can successfully predict more-aggregate phenomena and that can continually be tested with new empirical analysis are what ultimately generate scientific consensus. More theory building is needed in educational research.

Time on task still appears to be a central organizing concept in learning. A secondary concept involves the productivity and optimal division of that time among the different alternatives: presentation of new material through lectures, supervised and unsupervised practice, periodic repetition and review, and testing. Students have a wide variance in the ways they spend time in school. Part of the variance appears to depend on teacher characteristics, characteristics of other students in the class, and the amounts of time parents spend at home instructing children. Theories of learning need to be developed that incorporate school and home time and the various trade-offs and differences that exist across teachers, classrooms, and SES levels. Such a theory would generate a number of testable hypotheses for research that would then allow better and probably more-complex theories to be developed. Such theories can then provide guidance about which research is important to undertake.

The differences in effects between low- and high-SES students are particularly important to understand. One reason for this is that resource substitutions can occur between families and schools that can affect achievement. High family resources may often substitute for and supplement school resources in indirect and unmeasured ways that affect the accurate measurement of policy variables. Parental time spent on homework may substitute for individual

teacher time in the classroom, allowing the teacher of higher SES students to spend more time lecturing and thus avoiding the opportunity costs of individualized instruction inside the classroom.

Families may also shift more resources of time and money when school resources are lowered, and less when schools are devoting more resources to students. Thus, students with higher levels of family resources will be more immune to changing school resources than students with lower levels of family resources. This could help explain the weaker schooling effects for students in higher-resource families. Students from families with few resources show the most sensitivity to levels of school resources because the substitution potential is weak or nonexistent. However, the results of this analysis would imply that more school resources could substitute for lower family resources. These substitutions need to be the focus of much more research.

Improving NAEP Data

If NAEP would collect a **school district** sample rather than a **school** sample, historical data from school districts (not available at the school level of aggregation) and Census data could be used to obtain decidedly superior family and schooling variables for models. Census data can provide good family characteristics for school districts but not generally for schools. The necessity of including variables since school entry in specifications makes district-level samples necessary for developing analytical models below the state level of aggregation.

One additional advantage of moving to a district sample is that comparison of scores could be made for major urban and suburban school districts. Urban school systems pose the greatest challenge to improving student achievement, and being able to develop models of NAEP scores across the major urban school districts could provide critical information in evaluating effective policies across urban districts. The samples would be much larger than at the state level and could be expected to provide more-reliable results than for states.

If NAEP does not move toward a district-level sample, collecting a very limited set of data from parents should be considered. The critical parental information could be obtained with no more than 10 questions.

LIMITATIONS AND CAUTION

No single analysis of achievement scores is definitive. Rather, the coherent pattern that emerges across experimental and nonexperimental measurements and the associated theories that explain the mechanisms causing achievement gains in classrooms ultimately build scientific consensus and confidence in informing policymaking. We are still far from achieving this kind of consensus for the effects of educational policy variables. Until this happens, there will always be legitimate differences of opinion about the importance and interpretations of any empirical results.

We believe that providing a new explanation that more successfully encompasses the pattern of previous nonexperimental results, the Tennessee experimental data, the pattern of score gains in the 1970s and 1980s, and the new results in this report may be the most important part of this report for policy. While the results of the analysis of state scores can be important, developing an explanation that accounts for a much wider set of results—in the absence of competing explanations—is more important for policy purposes. However, competing explanations need to be proposed, and more research is needed that can test this explanation.

Finally, achievement scores are not the only, and arguably may not be the most important, output of schools. It may be possible to have good schools that are responsive to students and parents that do not place strong emphasis on achievement scores. It is certainly necessary to collect wider measures than achievement when assessing schools.

Although NAEP strives to reflect a broad range of items, so that some items reflect skills learned at earlier grades and some at later grades, the scores can reflect the timing of when students learn skills. Students in different states do not learn particular skills in the same sequence or at the same grade level. The types of state assessments done and whether these assessments are more or less similar to NAEP tests may also influence scores. States that have assessment systems that are similar to NAEP might be expected to score higher because of the alignment of curriculum with NAEP items.

The effects measured should be seen primarily as long-term effects of differences in characteristics. States should not necessarily expect

to see the full effects measured here in the first few years. The state differences measured here have, for the most part, existed over long periods, allowing students, teachers, parents, and curriculum to make longer-term adjustments.

“Teaching to the test” is often cited as a concern in assessments. Such a term carries three connotations. One connotation is a temporary inflation of achievement. Teachers are doing something that can result in a short-term achievement gain only, but the student’s achievement will not benefit in the long term. In this case, achievement scores can be misleading indicators, and testing can provide perverse incentives. A second connotation of “teaching to the test” is more positive and suggests that tests reflect accepted standards for what children should know and that review and repetition are necessary to achieve both short- and long-term gains in achievement. This possibility should be of less, if any, concern. A third connotation is that an imbalance occurs in the time spent on and the priority of tested versus untested subjects, or between educational goals related to achievement and those not related directly to achievement. If achievement gains occur at the expense of untested subjects or other socially desired objectives, some concern is warranted. In this case, broader measures are needed, and priorities should be set across objectives.

These concerns are more prevalent for “high stakes” tests, those for which there are consequences for students, teachers, or administrators. These concerns are minor for the NAEP, since no feedback or consequences are provided to students or teachers for NAEP tests. However, high-stakes state assessments could certainly be reflected in NAEP assessments to the extent that the tests are similar.