

This report presents results of a study that examined state-level achievement scores on the National Assessment of Educational Progress (NAEP) tests given in math and reading from 1990 through 1996. The report develops three measures that compare state performance: raw achievement scores, estimates of score differences for students with similar family characteristics, and estimated improvement trends. The analysis also focuses on measuring the effects on achievement of different levels of per-pupil expenditures and different policies that have significant resource implications and that have commonly been used in previous studies to explain achievement. The analysis also provides estimates of the cost-effectiveness of these resource-intensive policies. Finally, the report addresses whether there is evidence of score gains outside of resource-intensive variables that might indicate that diverse reform policies that have been widely implemented across states are raising achievement. The study represents a first step in understanding how various state policies, patterns of resource allocation, and reforms affect student outcomes and suggests directions for future research.

BACKGROUND AND MOTIVATION

States have always had significant influence over K–12 educational policies. That influence has increased even more during the latest wave of educational reform, dating from the mid-1980s. A broad and diverse range of new initiatives has been implemented, mainly at the state level. The initiatives include “systemic reform” efforts that establish and align standards with assessment, professional development, and some form of accountability for schools. Other initia-

tives include tightening certification and recertification standards for teachers, enhancing early education by subsidizing prekindergarten for lower-income families, and reducing class sizes in early grades. Many states also passed legislation authorizing charter schools, school choice, or contract schools. If reform policies are effective, these effects should appear in achievement gains and variations in trends in achievement gains across states.

States have a surprisingly wide degree of variation in the level of per-pupil expenditures and how they are utilized. Wide variation across states appears in nearly all educational measures, including

- Teacher-pupil ratios. In 1993, average pupil-teacher ratios for regular students varied from over 25 in California and Utah to under 15 in New Jersey, Vermont, and Connecticut.
- Spending per student. Levels of spending per student (adjusted for cost-of-living differences) varied from \$9,000 in New Jersey and New York to \$4,000 in Utah and Mississippi.
- Average teacher salary levels. Adjusted for cost-of-living differences, salaries ranged from over \$40,000 in New York and Massachusetts to less than \$30,000 in Mississippi.
- Teacher experience. The proportion of teachers with more than 20 years of experience varied from 11 percent in West Virginia to over 35 percent in Michigan, Pennsylvania, and Connecticut.
- Advanced degrees. The proportion of teachers with advanced degrees varied from over 80 percent in Indiana to less than 20 percent in North Dakota, Wyoming, and Idaho.

Such large variation in characteristics across states would offer a potential opportunity to measure their effectiveness if comparable measures of educational performance existed across states. Having 50 states take different approaches to education can provide a powerful advantage in the long run if research and evaluation can identify what works and what does not. Successful policies and practices can be adapted across states in a continual and ongoing process of improving education. Evaluating the effects of different and changing state policies then becomes an integral part of improving our schools and student outcomes.

Another reason to focus on states is that previous measurements of the effects of educational resources show quite different results if the

measurements are done at the state level rather than the district, school, classroom, or individual level. Measurements at the state level have shown very consistent and robust positive effects of added resources on educational outcomes, while measurements at lower levels of aggregation show less-positive and more-inconsistent effects. The debate about the effectiveness of educational resources has primarily used measurements at lower levels of aggregation. More of such measurements are available, and researchers have presumed that less-biased measurements occur at lower levels of aggregation. The consistent and robust positive state-level measurements are generally viewed as biased upward.

The inconsistency of measurements at lower levels of aggregation has provided major support for a view that public education has been ineffective and inefficient in using additional resources. This inefficiency is hypothesized to be a result of poor incentives within the bureaucratic public school system, which is seen as “unreformable.” In this view, providing more money to public schools is inefficient. A major focus of such reform efforts has been the attempt to circumvent existing public school structures by creating competition within the system or alternatives outside the system, including vouchers, charter schools, and outsource contracting for schools. A more-comprehensive assessment of student performance across states can help inform this debate by measuring whether different levels and allocations of resources across states affect achievement and the cost-effectiveness of various policy options. The results can also highlight states with different levels of unexplained performance on various measures related to achievement, thereby allowing more-intensive case studies to discover the source of these differences—particularly whether reform efforts are responsible.

STUDY OBJECTIVES

This study attempts to address these issues and has several specific objectives:

- to compare raw achievement scores across states and to determine which states have statistically significant improvements, taking account of all NAEP tests between 1990 and 1996

- to estimate NAEP scores for students with similar family characteristics across states to develop a better measure for the overall effects of educational policies and environments
- to determine whether trends and differences in scores across states for students from similar family backgrounds can be statistically linked to differences in state educational system characteristics that are resource intensive (including per-pupil expenditures, pupil-teacher ratios, public prekindergarten participation rates, teacher-reported adequacy of resources for teaching, teacher salary levels, teacher education, and teacher experience)
- to determine whether significant trends exist that are unaccounted for by these resource-intensive variables that might suggest effects from unobserved variables linked to reform efforts
- to estimate the costs of changing these resource-intensive policies and characteristics and to compare their cost-effectiveness in improving scores
- to propose a broader explanation for the pattern of achievement results reported here and in the empirical literature that also incorporates the new experimental class-size results and the historical pattern of spending and achievement in the nation
- to identify possible improvements in NAEP state data collection.

Given our results, we propose a broader explanation concerning the effectiveness of resources in the public school system that attempts to assess the pattern of previous nonexperimental results, the new results from experimental data, and the pattern of national score trends and resource growth from 1970 through 1996. This explanation states that additional resources provided to public schools mainly affect minority and less-advantaged students and that these effects can be large and significant if properly allocated and targeted. However, additional resources deployed in historical ways have had much less, if any, effect on more-advantaged students.

METHODOLOGY

Several issues have confounded attempts to assess student performance across states. Primarily, there have been no statistically valid measures of the achievement of representative samples of students

across states until recently. While many states have collected achievement scores within their states for a number of years, these scores are not comparable across states. This absence severely restricted the type and likely success of evaluations of state policies. One result is that state-level research has focused on collecting data on what states are doing and how they are doing it, but rarely on how different state policies and practices affect educational outcomes.

Comparative state analysis became possible when the Department of Education gave the NAEP tests to representative samples of students across a voluntary sample of states in 1990, 1992, 1994, and 1996. Seven tests were given in reading and mathematics at either the 4th- or 8th-grade level. Each test was administered to approximately 2,500 students, with 44 states represented in the sample. These tests represent the first valid, comparable measures of achievement of representative samples of children in various states.

While these tests presented an opportunity to answer these questions, there were significant barriers to carrying out analysis with these data and obtaining the kind of reliable results policymakers need. First, previous research suggests that family variables would account for a substantial part of the variation of scores across states because of the wide variation in their demographic composition and family characteristics. The family variables collected with NAEP were limited, and those collected were reported by 4th- and 8th-grade students, making their quality problematic. Without accurate family control variables, the effects of school resource variables would be biased upward, since family characteristics are positively correlated with schooling characteristics. The analysis needed to address this issue.

The second issue is that the sample was small; the state scores lacked independence across tests; and states participated in an unequal number of tests. Our sample represented 44 states, with a total of 271 scores. Most states in the sample took either six or all seven tests, but some took only two. The scores from the same states are not independent, which effectively reduces the sample further. Results from small samples can be more vulnerable to statistical assumptions, estimation procedures, and the influence of a few extreme data points, so the analysis had to test the sensitivity of the results to these conditions.

The third issue is the credibility of results derived from models aggregated across states. Unlike the generally null effects previously measured at lower levels of aggregation, previous studies using state-level data have shown that educational resources have consistent positive, statistically significant effects on educational outcomes. The interpretation of this disagreement has generally been that measurements using less-aggregate data are more accurate and that state-level results are biased upward. So, an alternative explanation is required for this discrepancy to make state-level results credible.

The fourth issue also involves credibility. Models using nonexperimental data will be deemed more credible if they can predict results that agree with results using experimental data. The Tennessee Student-Teacher Achievement Ratio (STAR) class-size experiment showed that reducing class sizes in K–3 had positive and statistically significant effects through 8th grade. The effects are generally larger for minority and disadvantaged students. A more recent quasi-experiment with pupil-teacher reductions in Wisconsin also showed initial results similar to those of the Tennessee experiment. Models using nonexperimental data therefore need to test predictions against these results.

We attempted to address these issues in our study. First, instead of relying on NAEP-reported family variables, we used Census data and data from the National Educational Longitudinal Survey—the largest survey collecting both achievement scores *and parent-reported family characteristics*—to develop three sets of family variables that use different sources of data and methods of weighting the influence of family characteristics. We estimated both fixed- and random-effect models that make different, but plausible, assumptions about the statistical properties of the data. We used random-effect models with the general linear estimator with exchangeable correlation structure to address the issues of unequal variance and number of observations across states and the lack of independence of observations. We performed a variety of sensitivity analyses to determine how sensitive the results were to extreme data points and to alternative statistical estimation and modeling procedures.

Second, we used a model specification consistent with the results from the Tennessee class-size experiment. The experimental results from Tennessee seem robust to the inevitable flaws that occurred in the implementation of the experiment, and these results provide

important lessons for specification of models with nonexperimental data, as well as important evidence about the effects of resources. The results of this experiment seem to indicate that including variables accounting for educational characteristics since school entry is important and that use of models incorporating pretests may be untenable. We also used our models to estimate a class-size effect for Tennessee and compared the results to the experimentally determined results from Tennessee. The results show agreement.

MAIN FINDINGS

Highlights of the Findings

Overall, the results paint a more positive picture of American public education than is commonly portrayed, especially with respect to effective allocation of resources. The following are some highlights of the findings:

- Public elementary students across states in our sample showed statistically significant gains (about 1 percentile point) in mathematics between 1990 and 1996.¹
- Some states are making significantly more progress than others. The math gains across states showed that a few made gains of around 2 percentile points a year, while others had almost no gains.
- The group of more-rural northern states had the highest average achievement scores, and southern states were usually among the lowest. The more-urban northern states generally fell closer to the middle of the score distribution. This distribution is explained mainly by family rather than school characteristics.
- There were statistically significant differences—as large as 11 to 12 percentile points—among students with similar family characteristics across states. All regions of the country had states with both higher and lower student scores from similar families.
- Both the level of expenditure per pupil and, more importantly, its allocation affected student achievement—particularly for states

¹The reading data are insufficient for analysis until the 1998 state NAEP reading data are included.

with disproportionately higher numbers of minority and less-advantaged students.

- Some educational expenditures are much more cost-effective than others. The difference in cost-effectiveness depends on how the expenditures are directed but can also vary markedly, depending on the SES level of the state, the current allocation of expenditures, and the grades targeted.

Evidence for the Effects of Reform

This analysis provides strong evidence that math scores from 1990 through 1996—controlling for population changes and participation rates—increased in most states for public school students by statistically significant amounts. Eighth-grade math scores increased more than 4th-grade scores. These math gains, which averaged about 1 percentile point a year, were far above the average gains experienced from 1973 through 1990. The small changes in resource-intensive variables during this period explain little of the improvement, so reform efforts would be the leading candidate to explain these gains. However, additional research, including case studies across states, is necessary to test adequately whether and which reform efforts may be linked to achievement gains.

Trends in reading scores cannot be assessed with the current data, since only two reading tests, given only two years apart, were available. The addition of the 1998 4th-grade reading test will provide a better assessment of national and state improvements in reading.

Some states had estimated math gains of approximately 2 percentile points per year, while some had little gain (see p. 62). Texas and North Carolina were among several states that made large, statistically significant gains, and state-administered tests also showed large gains during this period. The resource-intensive variables included in our analysis do not explain much of these gains over time. Therefore, reform efforts would be the leading candidates to explain the gains in these states.

Scores for Students from Similar Backgrounds

The scores of students with similar family and demographic characteristics varied by approximately 11 to 12 percentile points. Our

analysis distinguishes three groups of states: those whose scores for students from similar families are significantly above the median state, those whose scores are below, and a broad middle group (see p.68). Adjoining states and states with similar family characteristics often have statistically significant differences for students with similar family characteristics.

In part, these score differences can be traced to several systemic features:

- lower pupil-teacher ratios
- higher public prekindergarten participation
- lower teacher turnover
- higher levels of teacher-reported adequacy of resources for teaching.

Texas was in the highest group of states and California in the lowest on scores for students from similar families. The difference is about 0.34 standard deviations, indicating that similar students in the two states would emerge from K–12 education with score differences of about 11 percentile points. The variables in our model explain two-thirds of the difference in these scores. The major contributions to the higher Texas scores are lower pupil-teacher ratios, a much larger percentage of children in public prekindergarten, and teachers who have more resources necessary to teach. However, in-depth analysis using these measures as guides will be necessary to reveal the more complex of the features of state educational systems that create differences.

The Effects and Cost-Effectiveness of Educational Resource Allocation

Other things being equal, NAEP scores are higher in states that have

- higher per-pupil expenditures
- lower pupil-teacher ratio in lower grades
- higher percentages of teachers reporting adequate resources for teaching

- more children in public prekindergarten programs
- lower teacher turnover.

Other things being equal, states with higher teacher salaries or a higher percentage of teachers with master's degrees do not have higher scores. The lack of effect from direct investment in salaries from this analysis may have four explanations, and further research should be undertaken to help identify the reason. One explanation is that interstate differences in salary may be less sensitive to achievement than are intrastate salary differences. The primary teacher labor markets may be within states in which interdistrict salary differentials may affect the supply and distribution of higher-quality teachers much more than do interstate differences. A similar analysis across school districts within a state might show stronger compensation effects.

A second explanation is that teacher salary is the schooling characteristic that correlates most highly with family SES variables, and part of the salary effect may appear as social capital. If teachers teach children who have SES levels similar to their own, it may be difficult to separate salary and social-capital effects. The other variables in our analysis show much less correlation with family characteristics.

A third explanation is that these measurements occurred during a period of an adequate supply—even surplus—of teachers across most regions and types of teachers. Lower salary sensitivity would be expected when supply is more readily available. However, labor-market conditions are changing markedly for teachers because of demand increases from rising retirements, lower class sizes, and rising attrition rates, partly because of a strong economy. The supply of teachers is not expanding much, because the job market outside teaching is strong. Sensitivity to teacher salaries would be expected to increase under these conditions.

Finally, the results could partly reflect the inefficient structure of the current teacher-compensation system. The current system rewards experience and education—but neither seems to be strongly related to producing higher achievement. 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, in the current system, another dollar of compensation is used to reward experience and degrees and to raise all salaries—rewarding both high- and low-quality teachers—and teachers of both low- and high-scoring students. With such a compensation system, lower effects might be expected.

The effects of factors that influence achievement can vary markedly, depending on the type of students targeted and current program funding levels. For instance, lowering pupil-teacher ratios in states with high SES levels and current levels below the national average appears to have little effect. However, lowering pupil-teacher ratios for students in lower grades in states with low SES that have ratios above the national average has very large predicted effects. Prekindergarten also has much stronger effects in states with lower SES, while the adequacy of teacher resources appears to have significant effects for states regardless of family characteristics.

Taking into account both the costs and effects of policies, we found that the cost-effectiveness of resource expenditures could change by more than a factor of 25, depending on the program or policy, which types of students and grades are targeted, and the current program levels. The policies this analysis predicted to be most cost-effective include the following:

- providing teachers with more discretionary resources across all states
- in states with a disproportionate percentage of lower-SES students, lowering pupil-teacher ratios in the lower grades to below the national averages, expanding public prekindergarten, and providing teachers additional resources
- lowering pupil-teacher ratios in the lower grades to the national averages in states with average SES characteristics.

We also estimate that the use of in-classroom teacher aides is far less cost-effective than the policies cited above.

This analysis suggests that investing in better working conditions for teachers to make them more productive (lower pupil-teacher ratios, more discretionary resources, and improved readiness for school from prekindergarten) could produce significant gains in achieve-

ment scores. Conversely, 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.

THE BIGGER PICTURE: UNDERSTANDING EFFECTS OF INVESTMENTS IN PUBLIC SCHOOLS

Any general theory about the effects of public-school expenditures has to account for the following:

- the pattern of results in previous nonexperimental measurements
- the results of the Tennessee experiment and the Wisconsin quasi-experiment
- the pattern of national score gains and expenditure growth from 1970 through 1996.

One frequently advanced explanation holds that public schools have not demonstrated a consistent ability to use additional resources to improve educational outcomes. This explanation depends mainly on the inconsistency in nonexperimental measurements at levels of aggregation below the state level. It assumes that the inconsistency in measurements reflects inconsistency in the utilization of schooling resources rather than inconsistency in the measurement process. However, this explanation is not consistent with the experimental results from Tennessee or Wisconsin, with the large score gains for minority and disadvantaged students in the 1970s and 1980s, or with the positive and consistent nonexperimental results at the state level of aggregation.

We propose a different explanation that appears more consistent with the current experimental and nonexperimental evidence and historical expenditure and achievement trends. In our view, additional resources have been effective for minority and disadvantaged students, but resources directed toward more-advantaged students—the majority of students—have had only small, if any, effects. This explanation is consistent with the pattern of national score gains and expenditures from 1970 through 1996. Minority and lower-SES white students made significant score gains in the 1970s

and 1980s, but more-advantaged students made much smaller, if any, gains. These gains for minority and lower-SES students followed the national effort to focus on inequalities in educational opportunity and southern desegregation, and modest levels of additional real resources were provided in the form of compensatory programs, reductions in pupil-teacher ratios, and more experienced and educated teachers. National pupil-teacher ratios declined during this period, and evidence from our model and from the Tennessee experiment would suggest that such reductions are consistent with explaining part of the black student gains in the 1970s and 1980s.

The results of the Tennessee experiment and Wisconsin quasi-experiment show positive, statistically significant long-term effects on achievement. The samples for these experiments were disproportionately drawn from the minority and disadvantaged student populations. Our state-level results also produced estimates for pupil-teacher ratio that are consistent with the size of the effects measured in the Tennessee experiment and also produced a similar pattern of larger effects for minority and lower-SES students found in the Tennessee experiment. This agreement suggests that aggregate-level measurements may provide more unbiased effects than less-aggregate models.

Our explanation cannot account for the lower, and inconsistent, pattern of previous measurements at levels of aggregation below the state level. Most independent literature reviews now conclude that the previous nonexperimental results show that the effects of additional resources on educational outcomes are generally positive. But these reviews have not yet explained the wide variance in previous results or why more-aggregate measurements show more positive and consistent effects than measurements at lower levels of aggregation. We hypothesize that the inconsistency reflects the measurement process itself rather than inconsistency in the use of resources.

Previous measurements used widely different specifications and assumptions that may account for the inconsistency. Previous measurements also did not focus much on measuring separate effects for high- and low-SES students. If most measurements contained typical student populations with large proportions of more-advantaged students, smaller effects might be expected, and effects would be “inconsistent” across studies if student characteristics changed.

Effects may also differ across grade levels, leading to “inconsistent” results across studies that focus on measuring different grade levels.

We also hypothesize that measurements at lower levels of aggregation made with available previous data sets may be biased downward. The Tennessee experimental data identified one source of such bias: missing variables for years of schooling since entry. Another is the use of pretest scores as controls in production-function specifications. However, other forms of bias, including selection effects and differential quality and specification of family variables across levels of aggregation, may plausibly introduce differential bias at lower levels of aggregation. Further research is needed that focuses on the direction and magnitude of differential bias across levels of aggregation. If the source of inconsistency in previous measurements at lower levels of aggregation can be found, whether from bias or different specifications or student characteristics, a broadly consistent picture could emerge of the effect of resources on educational outcomes.

IMPLICATIONS FOR POLICY: IMPROVING AMERICAN EDUCATION

As we have noted, one interpretation of the empirical evidence implies that, in the absence of fundamental reforms of incentives and organizational culture, additional resources for public education are not the answer to improving schools. Underlying this view is the idea 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 foster greater competition for public schools.

Our results show that resources can make significant differences for minority and lower-SES students in public schools and that between-state, rather than within-state, differences in resources are the main reason for inequitable resource levels for lower-SES students. Such between-state differences can only be addressed with federal programs. However, our results also suggest that significant gains that cannot be traced to changing resources are occurring in math scores across most states. Although much research is required to attribute these gains to specific reforms, a plausible explanation would suggest that ongoing systemic structural reform **within** public education

might be responsible. These reforms may be linked to changing culture and/or incentives in public education or many other factors. But these results certainly challenge the traditional view of public education as “unreformable.” Public education may be a unique type of public institution that can achieve significant reform because it consists of a large number of separate, but diverse, units whose output can be measured and compared, leading to the identification and diffusion of successful initiatives. But some caution is warranted until these student gains in elementary schools result in longer-term gains in secondary schools and lead to completion of more years of education and to greater success in the labor market.

There are reasons to believe that improvements in achievement may 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. A significant contribution may also come from improving educational research and development by relying more on experimentation, focusing on improving the methodologies and assumptions inherent in nonexperimental data, and pursuing a coherent research strategy focused on using experimental and nonexperimental results to build successful theories of educational processes within families and classrooms.

IMPLICATIONS FOR RESEARCH

Experimentation and Improving Nonexperimental Analysis

Expanded experimentation in education is critical to understanding educational processes and helping to determine the appropriate assumptions and specifications to use with nonexperimental data. Experimentation should be directed both toward measuring the effects of major resource variables and toward the critical assumptions used in nonexperimental analysis. In addition, research—both experimental and nonexperimental—is needed that seeks an understanding of what changes inside classrooms and in student development when resources changed. Research consensus is unlikely to emerge until we understand what causes the differences in experimental and nonexperimental measurements and the differences among nonexperimental measurements and until we have theories explaining how changing resource levels affect parent, teacher, and

student behavior in the classroom and families and how these changes affect long-term student development in ways that result in higher long-term achievement.

Two hypotheses that arose from this analysis also need much more study. The first is the dynamic nature of achievement effects across grades, which the Tennessee experiment suggested. Schooling variables in one grade appear to be able to influence achievement at *all* later grades, so conditions during all previous years of schooling need to be present in specifications. It also appears that pretest scores may not adequately control for previous schooling characteristics. *The Tennessee results suggest that two students can have similar pretest scores and similar schooling conditions during a grade and still emerge with different posttest grades that have been influenced by different earlier schooling conditions.* For instance, despite having similar schooling conditions in grades 4 through 8, relative changes in achievement occurred in grades 4 through 8 for students having one to two or three to four years in small classes in K–3. Thus, the answer to the question of whether a smaller class size in 2nd grade had an effect cannot be known until later grades, and that answer will depend on what the class sizes were in previous and higher grades.

Conceptually, this makes the effect of class-size reductions resemble a human “capital” input that can change outputs over all future periods, and models that specify the effects of capital investments may be more appropriate. From the standpoint of child development, these results are consistent with the concepts of risk and resiliency in children. Children may carry different levels of risk and resiliency into a given grade that appear to interact with the schooling conditions in that grade to produce gains or losses. For instance, four years of small classes appear to provide resiliency against later larger class sizes, whereas one or two years do not.

A second key hypothesis underlying this analysis 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. Families may apply more of their own resources of time and money when school resources are lowered but apply less when schools are devoting more resources to students. Thus, students with higher levels of family

resources may be more immune to changing school resources than are 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. However, the results of this analysis would imply that more school resources can 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 into 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 more scores could be compared for major urban school districts. The urban school systems pose a large challenge to improving student achievement, and being able to develop models of NAEP scores across the major urban school districts could provide critical information for evaluating effective policies across urban districts. The sample sizes 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 ten questions.

ASSUMPTIONS AND CAVEATS FOR INTERPRETING THE STUDY RESULTS

Achievement is only one of many desirable outcomes expected from our schools. Until other comparable measures of outcomes are

available, test scores probably will receive a disproportionate share of attention. It is certainly possible to overemphasize achievement at the expense of other outcomes. It is also possible to have good schools that satisfy parents that may not be among the highest achieving. However, achievement is one important outcome expected of schools, and we should try to understand the policies that contribute cost-effectively to increase achievement and, at the same time, begin collecting a broader range of measures of school outcomes to achieve balance.

No test is a perfect indicator of what students have learned. Achievement scores reflect particular test items, and these items can emphasize more basic skills than critical-thinking skills. The NAEP state tests were redesigned in 1990 and reflect a mix of items testing more basic skills and some more-advanced, critical-thinking skills. Composite scores can mask important differences in kinds of achievement and knowledge, and more detailed analysis of sub-groups of questions is certainly needed to explore these differences.

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. Because of differences in curricula, 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 standards and assessment systems that reflect NAEP might be expected to score higher because the curriculum is aligned with NAEP items.

“Teaching to the test” is often cited as a concern in assessments. Such a term carries three connotations. One is a temporary inflation of achievement: Teachers are doing something that can result in a short-term achievement gain, 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 a 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 connotation should be of less, if any, concern. A third connotation is that an

imbalance occurs in the time spent and priority placed on tested rather than untested subjects or on educational goals related to achievement rather than 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 students and teachers receive no feedback or consequences for NAEP tests. However, high-stakes state assessments could certainly be reflected in NAEP assessments to the extent that the tests are similar.

The effects measured should be seen primarily as long-term effects of differences in policies. States should not necessarily expect to see the full effects measured in the first few years. The state differences measured here have, for the most part, existed over long periods, allowing students, teachers, parents, and curricula to make longer-term adjustments.

Our estimated differences in scores for students from similar families can reflect a variety of factors both related and unrelated to the education system. We have identified several factors related to the characteristics of the state educational systems that do account for part of the differences. However, these factors explain less than one-half of the differences. The remaining variance can arise from unmeasured family characteristics; unmeasured characteristics of the educational system; characteristics of other social-support systems for families and children; or particular factors creating social capital in states, such as foundations. The estimates made here are a first step to identifying further the factors within each state that contribute to achievement.

The effects and rankings presented here all have ranges of uncertainty associated with them that need to be taken into account in using these results for policy guidance. The effectiveness of certain policies measured across states can also hide certain context-sensitive factors that can make a factor either more or less effective. Implementation of similar policies can differ across states and local

school districts; therefore, the particular effects predicted here may vary depending on the local or state context. The effects cited here represent estimates given the current contexts and implementation existing broadly across states.

Finally, these results are meant to identify effective policies and states whose students from similar backgrounds are performing at different levels. This information is a first step toward further identification of policies and practices that contribute to higher achievement and to understanding the reasons constraining broader implementation of successful policies.

The tendency for policymakers to blame or to take credit for these achievement results should be tempered by at least three factors. First, the achievement results from 1990 through 1996 can reflect policies and practices from the early 1980s through 1996. Eighth graders tested in 1990 entered school in 1982, and their scores reflect the quality of education throughout their schooling. The 1996 4th-grade scores reflect more-recent policies. Second, many of the reforms initiated since the mid-1980s require significant organizational adjustments and affect schools, teachers, and students only gradually. So, the full effects of policies initiated since the mid-1980s will not be reflected in these scores. Third, the research and development community in education has been unable to provide consensus results or pilot-tested policies and practices that could guide policymakers and educators to more effective practices. Without a critical mass of high-quality research, policymakers lack the key process required to improve education systematically. Without good research and development, progress in education or any other area will be slow, uncertain, and inefficient.