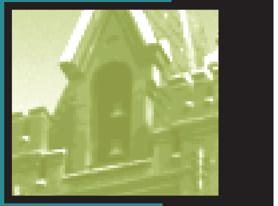
Analytic Issues in the Assessment of Student Achievement

Analytic Issues in the Assessment of Student Achievement



Proceedings from a Research Seminar Jointly Sponsored by

National Center for Education Statistics
National Institute on Student Achievement, Curriculum and Assessment
and
RAND

U.S. Department of Education

NCES 2000-050

The Charles Sumner School

More than any other school founded after the Civil War, the Charles Sumner School served as the cornerstone for the development of educational opportunities for black citizens in the District of Columbia. The significance assigned to its design and construction was indicated by the selection of Adolph Cluss as architect for the new building. In 1869, Cluss had completed the Benjamin Franklin School; in 1872, he completed Sumner School; and in 1873, he won a medal for "Progress in Education and School Architecture" for the City of Washington at the International Exposition in Vienna, Austria.

Dedicated on September 2, 1872, the new school was named in honor of United States Senator Charles Sumner of Massachusetts, who ranked alongside Abraham Lincoln and Thaddeus Stevens in leading the struggle for abolition, integration, and nondiscrimination. Upon opening, the Sumner building housed eight primary and grammar schools, as well as the executive offices of the Superintendent and Board of Trustees of the Colored Schools of Washington and Georgetown. The building also housed a secondary school, with the first high school graduation for black students held in 1877. The school also offered health clinics and adult education night classes.

A recipient of major national and local awards for excellence in restoration, Sumner School currently houses a museum, an archival library, and other cultural programs that focus on the history of public education in the District of Columbia.

Note: Cover photographs of the Charles Sumner School by David W. Grissmer

Analytic Issues in the Assessment of Student Achievement

David W. Grissmer
and
J. Michael Ross
Editors

Proceedings from a research seminar jointly sponsored by
National Center for Education Statistics
National Institute on Student Achievement, Curriculum, and Assessment
RAND

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
C. Kent McGuire
Assistant Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
1990 K Street, NW
Washington, DC 20006-5574

June 2000

The NCES World Wide Web Home Page is http://nces.ed.gov.

Suggested Citation: U.S. Department of Education. National Center for Education Statistics. *Analytic Issues in the Assessment of Student Achievement,* NCES 2000–050 by David W. Grissmer and J. Michael Ross. Washingon, DC: 2000.

For ordering information on this report, write:
U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794-1398

Contact:
David Grissmer
(703) 413–1100 ext. 5310
J. Michael Ross
(202) 502–7443

Or call toll free 1–877–4ED–Pubs.

Dedication

Dedicated to David W. Stevenson (1951–1998)

Senior Advisor to the Acting Deputy Secretary of Education, 1993–98

This book is dedicated to the memory of David W. Stevenson. His understanding of the interplay between basic research and education policy facilitated the development of this research seminar. From his early days in the sociology program at Yale, David began to develop a discipline-specific understanding of the structural factors mediating social change. As he became more involved in controversial policy issues, he saw the necessity for more definitive empirical evidence in their resolution. In the continual efforts of the research and policy communities, David's perspective will continue to enrich conversations about the direction of and appropriate methodologies for education reform. We acknowledge, with this dedication, his memorable accomplishments and our appreciation for his influence on this research seminar.

Table of Contents

SECTION I.
USING EXPERIMENTS AND STATE-LEVEL DATA TO ASSESS STUDENT ACHIEVEMENT

Section II. Using Longitudinal Data to Assess Student Achievement

Section III. Relating Family and Schooling Characteristics to Academic Achievement

Section IV. Policy Perspectives and Concluding Commentary

Section V. Appendix

Foreword

Peggy G. Carr
Associate Commissioner
Assessment Division
National Center for Education Statistics

In November 1998, a group of outstanding researchers and scholars gathered at the Charles Sumner School in Washington, DC to explore methodological issues related to the measurement of student achievement. Within this broad topic, the research seminar also focused more specifically on the sharing of perspectives related to the black-white test score gap. This sharing enabled the participants to compare their analyses and findings and to recommend improvements in data collection and analysis to the National Center for Education Statistics (NCES). Thus, eventually this collegial exchange promises to improve the utility of NCES data sets for policymakers in their efforts to ensure both excellence and equity in American education.

Seeking deeper explanations of the test score gap is a critical first step in the process of assessing student achievement more accurately. Toward that end, the seminar demonstrated the need for NCES to pursue more aggressively the development of concepts and methodologies that allow independent analysts to unravel the causes of such gaps. Such an “unraveling” requires closer examination of the complex interrelationships among resource factors, home and schooling influences, family configurations, and achievement outcomes. Further, NCES needs to place both cross-sectional and longitudinal data in a broader framework and to explicate our findings within diverse social contexts in richer detail.

The work of the Assessment Division in NCES, in particular, will benefit from the development of more explicit constructs that allow better comparisons of achievement results without the confounding interpretations that typically characterize conventional statistical presentations. For example, when achievement discrepancies between blacks and whites reveal different patterns in the northern states as compared to southern states, what type of analysis can we conduct that would enlighten our understanding of these historical and contemporary differences?

This first seminar has reminded us of the value of having researchers, scholars, and practitioners come together to advance knowledge in the field of achievement research and assessment. The collaboration of the sponsoring agencies—NCES, RAND, and the Office of Educational Research and Improvement (OERI) and its Achievement Institute—with their different missions, exemplifies the desire to integrate discipline-based perspectives toward common education reform goals. OERI and NCES acknowledge ongoing opportunities to sponsor a series of research seminars in order to ensure continued progress toward improving education policies and practices on behalf of our children and youth.

Seeking to engage a broader audience in this collegial exchange, NCES has prepared this volume containing the papers originally presented at the Charles Sumner School. The exchange of ideas among researchers and policymakers remains important to NCES. Still, this publication does not necessarily reflect the views of NCES or the policies of the U.S. Department of Education. Rather, the papers included here represent the views of their respective authors alone.

Acknowledgments

David W. Grissmer and
J. Michael Ross
Editors

The idea of a “research seminar” where academic researchers could share their current research findings with their federal counterparts took shape initially in early 1997. Ongoing discussions about the potential benefits of collaboration among the National Center for Education Statistics (NCES), RAND, and the National Institute on Student Achievement, Curriculum, and Assessment (NISACA) gave rise over the next year to a conceptual structure. A number of common interests were identified in the research and policy communities: periodic updates on complex survey designs and multilevel types of analysis. We went on to consider also our broader purposes: providing the direction to research that will inform policy developments in education, generating wider awareness of education research, and stimulating the development of better educational theory.

Within NCES, new forms of collaborative exchange were discussed. The one-day seminar received early support from Gary Phillips and Peggy Carr of the Assessment Division. Sharif Shakarani, then of the Assessment Division, helped to focus seminar offerings on NCES issues in data collection and analysis and fostered further collaboration by endorsing the participation of the different divisions in such a conference. Their understanding of the relevance of research updates shaped the concepts under discussion toward NCES needs. We are grateful, too, for Peggy’s strong and continuous advocacy and her financial support for the seminar. We appreciate also the substantive support offered by Holly Spurlock of the Assessment Division, whose careful and competent assistance throughout the process proved invaluable to the eventual success of the seminar. During this time, Daniel Kasprzyk, Director of the Schools and Staffing Program of NCES, also provided critical financial and moral support, and we remain grateful for his early commitment. The emerging plans for the seminar received support from Pascal D. Forgione, then Commissioner of NCES, whose sentiments were always directed toward pro-

viding the best research possible in the interests of assisting policymakers to improve education.

As the planning progressed, Joseph Conaty of NISACA provided ongoing insights in organizational support through his contacts with the academic research community. For the critical collaborations he contributed to this endeavor, we express our continuing appreciation. Further, we acknowledge the contributions of Marian Robinson, then an intern in Joseph's office, now at the Graduate School of Education, Harvard University, who smoothly executed numerous details of planning for the seminar.

We would also like to thank Marilyn McMillen, Chief Statistician for NCES, for broadening the base of participation in the seminar through the provision of special funding to cover the travel expenses of graduate students.

The role of the Education Statistics Services Institute (ESSI) in furthering the broad research and development purposes of the seminar is also very much appreciated. ESSI's ability to facilitate "making the seminar happen" made it possible for us to extend collaboration and consultation among all the participating groups. Another colleague whose support was critically important in the early stages of development is John Mullens, now of Mathematica Policy Research, who worked on the project under the auspices of ESSI. John offered substantive contributions to discussions about the importance and structure of the seminar, and then cheerfully took the lead in facilitating arrangements among all the parties. Later, he played an important role in ensuring that the early drafts of the solicited papers arrived in time for review before they were distributed to seminar participants. The benefits of the seminar were enhanced by John's grasp of the issues in research and policy and his facilitative skill.

Our appreciation for managing critical details extends to Bridget Bradley, then a consultant with Policy Studies Associates and later Policy Analyst in the Office of the Deputy Secretary of Education, who offered invaluable logistical support to our efforts to plan the seminar. Her gracious manner complemented her careful attention to making and monitoring arrangements, and we thank her sincerely for her efforts.

We extend *very* special thanks to the organizing committee that had major responsibilities for planning and staging the seminar, as follows: Peggy Carr, Holly Spurlock, and Daniel Kasprzyk, as well as John Mullens. We very much appreciate the committee's efforts, through the seemingly endless meetings, messages, and phone calls. Further, this committee, along with Brenda Turnbull of Policy Studies Associates and Martin Orland of the Early Childhood and International Crosscutting Studies Division (ECICSD), also participated in the detailed planning for the publication of the proceedings, and we are indebted to them for their useful suggestions regarding major decisions about this book. The benefits of their efforts on behalf of the seminar should be seen for years to come, as NCES endeavors to ensure continuous improvements in data quality and analytical methods.

On November 9, 1998 at the Charles Sumner School in Washington, DC, the seminar took place with approximately 100 participants in attendance. Titled "Analytic Issues in the Assessment of Student Achievement," the research seminar was jointly sponsored by NCES; the National Institute on Student Achievement, Curriculum, and Assessment; and RAND, as we had planned for so many months. The beautiful setting, the quality of the papers and the commentary, and the collaborative and collegial nature of the day's deliberations were the fruition of the long process of preparation.

With appreciation, we acknowledge the "silent" reviewers of the early drafts of the solicited research papers. Their early reviews increased the usefulness and applicability of the presentations and papers. These reviewers, in addition to the editors, were Martin Orland, John Ralph, Dan Kasprzyk, Peggy Carr, Joseph Conaty, and Holly Spurlock. Their work, though behind the scenes, was an important contribution to the substance of the seminar, and we appreciate their assiduous reviews.

Subsequently, the papers were forwarded to the colleagues who had agreed to serve as discussants for the seminar. Sylvia Johnson (Professor of Education at Howard University), Robert M. Hauser (Professor of Sociology at the University of Wisconsin-Madison), and Valerie E. Lee (Professor of Education at the University of Michigan) undertook the task of reviewing each pair of solicited research papers representing the methodological and conceptual strands of the seminar, seen here in Sections I, II, and III. Their comments enabled the authors of the solicited papers to make further improvements in their works before the seminar; then the discussants prepared their public responses for the presentations made during the seminar. We remain grateful for their dedication to this time-consuming task that benefited all seminar participants.

Similarly, we offer our appreciation to Marshall S. Smith and Christopher Jencks, whose presentations lifted our attention from such narrow topics as sampling design and dataset linkages to take a broader look at the effects of past analytical methods upon social scientists' understanding of achievement disparities and to share insights into how those understandings have played a role in the development of new education policies. Smith and Jencks, each in his own way and from his own perspective, explained the vagaries of education research since "the Coleman report" and went on to describe the usefulness of better data collection and analysis and of better theories and models.

Further, we acknowledge with appreciation the assistance of Joseph Conaty, John Ralph, and Martin Orland as moderators for the discussions during the seminar, as well as the participation of the seminar attendees (listed in the appendix), whose comments enriched the discussions and, therefore, the overall outcomes of the seminar.

Following the event, we made the decision to edit the proceedings for publication, recognizing the far-reaching implications of the discussions for NCES and desiring to extend the insights to a broader audience. Even more ambitious were our later decisions to include the Introduction and the fourth section, Policy Perspectives and Concluding Commentary. It was fortunate that Anne Meek of ESSI was available for the tasks that these decisions required. As a professional editor working closely with us, Anne ensured both the completion of the book and its internal coherence. We acknowledge with appreciation her grace and her sense of humor throughout the process of preparation.

In the preparation of this book, special thanks are due to Ron Miller of RAND for the design of the cover of the book (which incorporates a photograph by David Grissmer). We also acknowledge the assistance of staff at ESSI who prepared the proceedings for publication, as follows: Allison Arnold, Mariel Escudero, Anne Kotchek, Qiwu Liu, Jennie Romolo, and Jennifer Thompson. We thank them for their attention to detail and their technical skills, which have greatly improved this book for use by researchers, policymakers, and educators.

The persons named here have provided varied kinds and levels of support for the seminar and for the production of this book, and we are pleased to acknowledge our debt to each of them. However, the final responsibility for this publication rests with us, and any remaining deficiencies are solely our responsibility.

Introduction

Toward Heuristic Models of Student Outcomes and More Effective Policy Interventions

C. Kent McGuire
Assistant Secretary
Office of Educational Research and Improvement

In November 1998, in the research seminar commemorated here in this volume, a diverse community of scholars and researchers paused amidst their heavy schedules to turn their attention to a questioning of their methods of conducting empirical inquiries. Taking stock of a body of work is, of course, commendable for a professional group. It is always instructive to learn from one another and to consider how to better our efforts; and this seminar provided ample opportunity for such learning and consideration along several dimensions.

The seminar, however, went beyond the normal technical matters that education researchers typically discuss on such occasions. Rather, the gathering also shed light on research and policy issues, especially the continuing efforts to improve the performance of American education, to enhance greater educational equality of opportunity, and to understand the sources of continuing race-ethnicity achievement discrepancies. These larger purposes are, after all, the reasons we collect and analyze data in the first place and the reasons we search for improvement in our methods of data collection and analysis.

That the deliberations took place at the Charles Sumner School was especially appropriate for the Office of Educational Research and Improvement (OERI). Sumner School, now restored and an architectural treasure of great beauty, has long served as an important symbol of minority education. In this setting, we were surrounded by a particularly fitting sense of history for this discussion of both the means for measuring student achievement and the reasons for doing so.

The deliberations were enriched by multiple disciplinary perspectives. The research seminar included sociologists, economists, and education researchers, both new and more established researchers, and federal policymakers, all of whom shared their insights with each other. That is, researchers from different disciplines and methodological backgrounds commented on each other's analyses and listened to each other's recommendations, and federal policymakers provided their perspectives on the role of research and the important questions that must be addressed. In short, the seminar provided an enlightening forum for the exchange of perspectives and research findings, as participants contributed their particular expertise to discussions about the measurement of achievement and the contribution of education research to the improvement of schooling.

Of particular importance are some new insights in the understanding of racial and ethnic differences in student achievement. Such differences were first brought to our attention nearly 30 years ago by "the Coleman report," when the nation began to move *equality of educational opportunity* to its enduring place on the nation's agenda. Since then, we have come to understand much more about the variables associated with both high and low achievement—not nearly as much we would like to know but certainly more than we once knew. And OERI has always hoped to play a pivotal role in the empirical examination of these questions.

Over the past 10 to 20 years, the federal government has been improving its data collections, and a wide array of analyses continue to be conducted to move our understanding beyond Coleman's findings. These continued adjustments and processes have helped us to understand the complexity of what we are trying to measure and what we are trying to change. A brief synthesis of the papers solicited for this seminar will serve to illustrate the details of different data sets and, at the same time, help us to understand the systemic obstacles to changes in educational policies.

The papers are organized under three major divisions: (1) Using Experiments and State-level Data to Assess Student Achievement, (2) Using Longitudinal Data to Assess Student Achievement, and (3) Relating Family and Schooling Characteristics to Academic Achievement. The last major division, (4) Policy Perspectives and Concluding Commentary, presents important observations about research methodology and funding and the connections between research and policy, both with a retrospective view and a view toward the future.

Using Experiments and State-level Data to Assess Student Achievement

In the first essay, Stephen Raudenbush characterizes the state proficiency means from the Trial State Assessment of the National Assessment of Educational Progress (NAEP) as “difficult to interpret and misleading.” It is their multidimensionality that makes proficiency scores difficult to interpret: they may look simple at first glance, but actually they reflect many factors—student demographics, school organization and processes, and state policy influences. Raudenbush discusses his multilevel analyses that compare states on their provision of student resources for learning. Not surprisingly, he finds that socially disadvantaged students and ethnic minority students (particularly African American, Hispanic American, and Native American) are significantly less likely than other students to have access to advanced course-taking opportunities, favorable school climates, highly educated teachers, and cognitively stimulating classrooms. He also finds substantial variation across states in the extent of inequality in access to such resources. Such findings point, as he said, toward “sharply defined policy debates concerning ways to improve education.”

Grissmer and Flanagan speak from a different but equally illuminating perspective. Their major focus, fueled by concerns about inconsistency in research results, is the lack of consensus across the broad and multidisciplinary research communities in educational research. In many respects, of course, this lack of consensus has been inevitable, given the different research perspectives; the varied points of view expressed by researchers, policymakers, and practitioners; and the inherent complexity of education. Grissmer and Flanagan believe, therefore, that improvements in data collection and statistical methodologies, by themselves, are not sufficient to bring about the kind of consensus needed to effectively guide educational policies. Thirty years of research with nonexperimental data have led to almost no consensus on important policy issues, such as the effects of educational resources and educational policies on children and the impact of resources on educational outcomes. Further, they propose to guide the process of creating consensus through the development of a strategic plan, which would enable experimentation and data collection to provide the quality of data necessary for theory-building and also improve the specifications of models used in nonexperimental analysis.

Grissmer and Flanagan therefore recommend three approaches likely to lead to consensus: increasing experimentation, building theories of educational process, and improving nonexperimental analysis. They suggest that experiments have two main purposes: they provide the closest-to-causal explanations possible in the social sciences, and they help to validate model specifications for nonexperimental data. They present detailed discussions of important policy issues and the findings of research, including critical analyses of the "money doesn't matter" issue and the issue of the effects of resources on achievement, with examples from the many ways researchers have addressed these questions over the years. They also provide insight into such efforts as the Tennessee class size experiment, the use of NAEP scores and SAT scores, and new methods of analyzing education expenditures.

In addition to making some methodological recommendations, Grissmer and Flanagan explain the process of theory-building cogently and clearly. To advance theory-building, they advocate linking the disparate and isolated fields of research in education, for example, linking the micro-research on time, repetition, and review with the research on specific instructional techniques, homework, tutoring, class size, and teacher characteristics. Further, to enhance the development of modeling assumptions, they recommend linking the research on physical, emotional, and social development, differences in children, delays in development, and resiliency factors. Their suggestions for improvements encompass the need for experiments, improvements in NAEP data such as collecting additional variables from children, and supplemental data from teachers, among other things. All in all, their paper offers timely and thought-provoking views about the research community's next steps in improving theories of education and models of research, so that eventually the nation can indeed achieve its desired goals in education.

Using Longitudinal Data to Assess Student Achievement

Next, Meredith Phillips offers a number of convincing and far-reaching observations about improving methods of data collection and analysis, especially in efforts to understand ethnic differences in academic performance. Perhaps most relevant is her observation, echoed by other presenters, that we must study ethnic differences explicitly despite their political sensitivity. She explains that socioeconomic factors do not overlap with ethnicity as much as researchers have traditionally assumed. Ethnic differences in learning vary between the school year and the summer; therefore, the importance of collecting data in both spring and fall of each school year should be a major point of empirical queries. Further, since the test score gap widens more during elementary school than during high school, and children's test scores appear less stable during elementary school than during high school, Phillips also calls for focusing more surveys on elementary students rather than on high school students. Of particular interest is her assertion that we have learned little about ethnic differences because researchers have not adequately studied education outside of the formal institution of schooling. Measuring the cognitive skills of infants and toddlers prior to their entry into school could help to clarify ethnic differences in family influences on achievement. Phillips concludes by reminding us that "it is not logically necessary to understand the causes of a social problem before intervening successfully to fix it." To those who bear responsibility for the improvement of American education, this reminder is somewhat comforting, in view of the breadth and depth of recommendations made by this network of researchers and scholars.

Ferguson and Brown then discuss the relationship of teacher quality to student achievement, in particular, the relationship of teachers' certification test scores to students' test scores. The evidence they have assembled suggests that the black-white test score gap among students reflects a similar test score gap among teachers. From several studies, they cite findings suggesting that "teachers' test scores do help in predicting their students' achievement." For example, scores on the Texas Examination of Current Administrators and Teachers (TECAT) turned out to be strong predictors of higher student reading and math scores in school districts across the state. Ferguson and Brown explicitly make the point that ensuring well-qualified teachers in districts where minority students are heavily represented is "part of the unfinished business of equalizing educational opportunity." In Alabama, certification testing reduced entry into teaching by candidates with weak basic skills and consequently narrowed the skills gap between new black and white teachers. Since the rejected candidates would probably have taught disproportionately in black districts, Ferguson and Brown suggest that the policy of initial certification testing is probably helping to narrow the test score gap between black and white students in Alabama. Predictive validity has not yet been used as a criterion for validating such exams; still, Ferguson and Brown contend that policymakers can safely assume a positive causal relationship between students' and teachers' scores.

Relating Family and Schooling Characteristics to Academic Achievement

Brewer and Goldhaber offer additional insights into the relationship of student achievement and teacher qualifications, based on their analyses of data from the National Education Longitudinal Study of 1988 (NELS:88). Their linking of student-teacher-class elements in NELS:88 permitted these researchers to investigate the effects of specific class size, teacher characteristics, and peer effects on student achievement, through the use of multivariate statistical models. The NELS:88 data enabled the researchers to link students to their particular teachers and specific courses. In their analyses, they find that subject-specific teacher background in math and science is positively related to student achievement in those subjects, as compared to teachers with no advanced degrees or with degrees in non-math subjects. They did not see this pattern repeated in English and history. Nor did they find positive effects on achievement associated with teacher certification or years of teaching experience.

While encouraged by the recent improvements in data collection exemplified by NELS:88, Brewer and Goldhaber make pertinent recommendations for future data collections. Seeing the link between students and teachers as critical, they strongly recommend that such links not only be maintained, but also strengthened by the collection of additional data about teachers' backgrounds. Specifically, they suggest the addition of teacher test scores, the years that teachers obtained their licenses, and the states where they were licensed. Such data would be quite useful now and in the future, since policymakers in many states have recently overhauled or are considering changing licensure and/or teacher preparation requirements.

Brewer and Goldhaber point out that items relating to student, parent, and teacher beliefs, attitudes, and feelings could be omitted from data collections, since policymakers can only indirectly affect these. Further, they raise the questions of de-emphasizing the collection of nationally representative samples or of sampling fewer schools with more data on students and classes in a smaller number of schools. Brewer and Goldhaber are seeking the data quality necessary for the use of multivariate statistical models, because researchers find such models most persuasive in tackling important policy questions. Brewer and Goldhaber clearly state their belief that the "ultimate reason to collect data is to influence public policy in a positive way," a perspective that supports the continued improvement of data collection and methods of analysis.

Finally, in their investigation into school-level correlates of student achievement, McLaughlin and Drori report linking three sources of data: (1) data from the Schools and Staffing Survey (SASS) regarding such school and background factors as school size, class size, normative cohesion, teacher influence, student behavioral climate, teacher qualifications, and the like; (2) student achievement data from statewide assessments; and (3) data from the 1994 State NAEP fourth grade reading assessment in public schools. These researchers constructed a set of 18 composites of data on student background, organizational aspects, teachers' qualification, and school climate perceptions, then merged them with school reading and mathematics mean scores. McLaughlin and Drori analyzed the relationships of various school organizational factors to student achievement, hoping to elicit evidence on the correlations between school reform policies and achievement. An important finding is that reading scores were higher in schools with smaller class sizes. This finding was consistent across grade levels. Another interesting finding is that 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.

Perhaps more exciting than their findings, however, is the methodology McLaughlin and Drori employed and its potential for identifying effective school policies. Teasing out the correlates of student achievement through such linkages of databases is a promising venue for researchers and policymakers alike, especially since a number of states are turning to reforms that establish consequences for schools based on their gains in achievement over years.

Policy Perspectives and Concluding Commentary

Midway through the seminar, Marshall S. Smith engaged seminar participants in a retrospective look at past policy efforts to monitor and mitigate the discrepancies in black-white achievement scores. In his paper, he discusses possible explanations for the status of the gap at various points in time and concludes by reviewing current policy directions that promise further improvements in student achievement and recommending increased attention to experimental field trials.

Smith describes the reductions in the black-white achievement gap from 1971 through 1988, as seen in data from NAEP assessments, referring to a paper that he and Jennifer O'Day published in 1991, which reviewed policy initiatives and changes in student achievement 25 years after the Coleman report. Smith, who was at that time dean of the graduate school of education at Stanford University, pointed out in his presentation that these reductions reflected consistent and substantial increases in black scores and almost no change in white scores. In less than 20 years, the reduction in the achievement gap between black and white students was 33-50 percent in reading and 25–40 percent in mathematics, according to NAEP data.

Smith summarizes several tentative explanations for this reduction in the gap, which occurred between 1971 and 1988, which he and O'Day had first discussed in their paper. They had recognized, first, the large decrease in the percentage of black children living in poverty: from 65 percent in 1960 to 42 percent in 1980. Another highly plausible explanation was that preschool attendance increased substantially for low-income children. Further, Smith notes, the educational quality of schools for black students was dramatically enhanced with the dismantling of the old dual school system. In addition, the effects of Title I—while difficult to assess by numbers alone—included an increase in educational resources in schools, lower class sizes, and an emphasis on the basics of reading and mathematics. And, as Smith reiterated during the seminar, Title I also served to focus national attention on the needs of low-income students, many of whom were African American.

Smith reminds seminar participants that he considered the basic skills movement an influence in reducing the achievement gap at the secondary level during this period. After all, by the mid-1980s over 33 states had required students to pass a minimum competency test as a criterion for graduation. The resulting instructional emphasis on basic skills, combined with the "high stakes" tests, produced the focus and coherence in the curriculum needed for improving student achievement.

Smith goes on to speculate that, by 1990, the effects of the factors identified by him and O'Day had begun to diminish in their influence and that, therefore, the gap between black and white students' test scores was no longer continuing to narrow. Thus, the current task for policymakers has become to identify and implement policy ideas that promise to continue the process of reducing the gap initiated in prior decades. This task means thinking hard about, and also building upon, the interventions that brought about the earlier improvements in achievement.

Smith describes three major objectives at the federal level designed to support efforts to improve education in general and also to reduce the gap. The first is to create overall conditions as stable and livable as possible for all families with children. Smith cites, as efforts toward this objective, recent sustained economic growth and specific policies such as the Earned Income Tax Credit and the Children's Health Insurance Plan. The second objective is to expand educationally rich opportunities for all students beyond typical school schedules. As specific examples Smith lists the development of education standards for the Head Start curriculum, the expansion of Head Start enrollment, and increased services through the 21st Century After-School Program. The third is to encourage state and local standards-based reforms. Toward this end, federal programs such as Title I and Goals 2000 have been aligned to support the state reforms.

Standards-based reform, considered one of Smith's major contributions to education policy, in effect extends the basic skills movement to a much broader scope, with *all* children expected to attain the higher content and performance standards, not just basic skills. Even at such an early date as this, it is worth examining the promise of such reforms by looking at outcomes within the states. What have been the test score results in states with focused and coherent strategies in their standards-based reforms? Using NAEP data, Smith finds encouraging results in those states—especially North Carolina and Texas—with relatively challenging standards, curriculum-aligned tests, accountability provisions, extensive teacher training, and special efforts on behalf of low-scoring students. It is apparent that, for whatever reasons, some states are doing very well in their efforts to improve student outcomes, while others are not. Therefore, policymakers are obliged to consider very carefully the evidence about interventions that promise to lead to improved student performance.

Moving to a prospective view, and building a case for increasing experimental efforts, Smith cites the strength and authority of such studies as the Tennessee class size study and those on early reading acquisition at the National Institute of Child Health and Human Development (NICHD). He identifies several areas where policy development could well be more adequately informed through such studies; for example, methods of incorporating technology into classrooms, the effects of summer school, and replications of the NICHD studies. Smith argues eloquently for increasing the use of experimental field trials in education research and suggests that a list of recommendations for consideration for the research agenda at the Department of Education might come from the seminar.

Indeed, as Christopher Jencks pointed out in his presentation, for those who believe that educational policy should be based upon a more solid evidentiary structure, the current shortage of any type of randomized field trials in education policy represents perhaps the greatest challenge facing education policymakers and researchers alike. More pointedly, of course, OERI faces this challenge in designing a course for its own research agenda. According to Jencks, a major advantage of experimental studies is that the more persistent and difficult policy questions can be answered more definitively by the inclusion of randomization procedures at the school and classroom levels. These questions cannot be answered by improved data collections, more complex surveys, or more refined statistical methods alone. Critical policy questions such as the debate over ability grouping can be intensely controversial; and to resolve such questions by randomized field trials would still entail some unavoidable political fallout, no matter how definitive the findings.

Then, too, Jencks notes that the idea of randomized trials is rarely accepted within the field of education research. There are a number of practical obstacles to utilizing experimental methods: they inevitably change established school routines, since they necessarily include randomization of students or teachers to different schools or classes. It might be possible to convince educators that such procedures would constitute a small price to pay, given the very useful information to be gained, if only the researchers themselves strongly supported experimental studies. Jencks notes, however, that most education researchers are typically unenthusiastic about randomized experiments. In fact, he contends that most researchers now have limited knowledge of classic experimental studies.

Still, Jencks insists that the advantages of randomized field trials to policymakers are large and attractive. The first advantage lies in the knowledge to be gained from wider use of experimental methods; the second, in the clarity of understanding that results from these intuitively obvious methods. A legislator or a school board member, for example, can follow the logic of the Tennessee class size experiment, understand how the results were evaluated, and see why the results are consistent with what researchers say they mean. Nevertheless, Jencks is not suggesting that we abandon descriptive types of research proposals. On the contrary, surveys and experiments complement one another, each yielding valuable results necessary for providing the data necessary for policymaking. But the present dearth of experiments sounds a warning to OERI and highlights an imperative need for the next few years.

Indeed, with such different perspectives and challenging viewpoints brought to bear on a single topic, many possible directions were identified for the future work of NCES and OERI. Throughout the seminar, presenters and participants were persuasive in their descriptions of the necessity of complementing longitudinal survey data with data collected in the classical research design tradition such as the Tennessee class size experiment. Their praise for renewed consideration of experiments made this issue the predominant theme of the seminar, and one with far-reaching implications for the sponsors of the event.

Taking stock of our empirical methods—more or less the primary reason for organizing the seminar—yielded a second theme in the comments from presenters and participants. This theme was seen in the abundance of proposals for improvements in the design and analysis of data collections, including ways of making longitudinal studies more elaborate; suggestions about the addition or deletion of certain types of items on surveys; sampling more students per teacher; collecting longitudinal data more frequently; and gathering more measures of teacher quality. Implicit in many of the recommendations is the idea of more critical evaluation of the utility of variables and methods in all NCES surveys, whether longitudinal or cross-sectional, in order to design better surveys in the future. These suggestions translate into serious considerations for OERI and NCES as they move forward with new assessments of student achievement, as well as with all other surveys and analyses.

Last but by no means least, seminar participants emphasized the importance of communication among the different research disciplines. They referred specifically to the power of experiments to communicate effectively with policymakers and other researchers. They expressed appreciation for the seminar as a good example of such communication and recommended more such opportunities. The value of the seminar can easily be seen in the broad, data-based dialogue among researchers about the choices facing NCES and OERI and presented in this book. Suggestions were made to open the door to new partnerships among federal, state, and private researchers and to establish connections between state-based researchers and federal researchers. Interestingly enough, repeated references to the benefits to be gained from openness to a variety of audiences constituted a sub-theme of the seminar. Communication is, after all, an essential component of building consensus among researchers, scholars, and policymakers.

In short, the exchanges of this seminar promise researchers and policymakers alike that racial and ethnic differences in achievement can be explored more effectively than at present, that schools can continue to move toward equality of educational opportunity, and that progress toward the improvement of American education requires our continued communication, collaboration, and commitment. It is now our task to translate our knowledge into improved policies and practices in education for the benefit of our children and our nation.

Section I.
Using Experiments and State-level Data to Assess Student Achievement

Synthesizing Results from the NAEP Trial State Assessment

Stephen W. Raudenbush
School of Education and Survey Research Center
University of Michigan

During the past two decades, U.S. researchers, policymakers, and journalists have expressed concern that the nation’s schools are failing to prepare students to meet the demands of the modern global economy. Researchers have interpreted international assessments as revealing serious weaknesses in mathematics and science proficiency (see, for example, Beaton et al.1996; Medrich and Griffith 1992; NCES 1995, 230–231). Although such claims can be strongly contested (c.d., Rotberg 1998), they support a broader climate of malaise, and even crisis, concerning the performance of U.S. schools.

In this climate, calls for reform and accountability at every level of the education system have taken on greater urgency. The stakes are often high: students in Chicago must pass a citywide test to be promoted to the next grade; students in Michigan can obtain endorsed diplomas only by passing the state’s proficiency test; teachers with high-scoring classrooms can obtain cash rewards in some districts; and school principals are held accountable for school mean achievement.

For comparisons at the state level, the key source of data is the Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP), “the Nation’s Report Card” (c.f., Mullis et al. l992). Administered every two years (though in different subject areas at each administration), TSA enables cross-sectional comparisons among participating states in several subject areas at several grades and allows estimation of trends in student mean proficiency over time. Participation has grown to include more than 40 states and U.S. territories. But what are we to make of such comparisons between states?

Most “users” of the TSA would like to view state proficiency means as reflecting the effectiveness of educational provision, policy, and practice within each state. If so, TSA would provide direct evidence of the quality of each state's educational system. Talking to those involved in reform, for example, I have found it common to view California's performance on TSA in certain subject areas as direct evidence of the failure of reform in that state. Yet even a cursory examination of TSA data reveals that state demographic composition, including poverty levels and ethnic composition, is strongly associated with state mean proficiency—and state trends in proficiency are undoubtedly associated with state trends in demography. Thus, critics claim that state means are surrogates of demography more than indicators of educational effects. This criticism has led to many calls for statistical adjustment of state means on the basis of student social and ethnic background. Indeed, it is possible to compare states within strata defined by ethnic background and parental education (as in Mullis et al. 1992), but such within-stratum comparisons control background differences only roughly and do not take into account the extent to which a school's demographic composition creates a context affecting student performance.

The National Assessment Governing Board, which provides policy direction to NAEP, has resolutely rejected the notion of reporting statistically adjusted state mean proficiency. Board members fear that adjustments for student background will lower expectations for school systems serving disadvantaged students. There are also sound statistical reasons to be skeptical about adjustments. Suppose, for example, that we use a regression analysis to compute state mean residuals, that is, discrepancies between the actual state means and the means expected on the basis of student composition. Such residuals have often been interpreted as indicators of the "value added" by the schooling system. Yet, if the regression model fails to include key aspects of educational policy and practice, the estimates of the association between student composition and outcomes will be biased. The bias would arise because the quality of educational provision and student composition would be positively correlated, with the most advantaged students tending to be found in the schools with the most favorable resources, policies, and practices. Failing, then, to control for the quality of educational provision will inflate estimates of the contribution of student demography. This inflation, in turn, will lead to biased "value added" indicators. The result is an over-adjustment for demography, such that systems serving the most advantaged students will tend to look less effective than they are. However, the magnitude of the over-adjustment is impossible to assess in the absence of data on the quality of school policy and practice (see Raudenbush and Willms [1995] for a thorough discussion of this problem in the context of school evaluation).

Interpretation of state proficiency means is thus terribly risky. We cannot equate unadjusted state mean proficiency with educational effectiveness as many reformers wish, yet adjusted means set up low expectations for states serving poor students and are statistically untrustworthy.

The problem of interpreting the results of the TSA frames the pair of investigations I shall discuss in this paper.[1] The debate over the meaning of state mean proficiency reflects a longstanding debate about the sources of inequality in academic achievement in the United States. If inequality in family background is the key to inequality in educational outcomes, then inequality in aggregate family background ought to be key to understanding differences in state achievement means. On the other hand, if inequality in school quality is key to understanding inequality in individual outcomes, then aggregate school quality ought to explain state variation. Fortunately, NAEP provides some reasonable data at the level of both the student and the school to test these propositions.

Our first investigation, then, tested models for student math proficiency within each of the participating states of TSA. This may be likened to a “meta-analysis” in which each state’s data provide an independent study of the correlates of math proficiency. We examined student social, ethnic, and linguistic backgrounds, and home educational resources as predictors of student proficiency. Yet our models simultaneously included indicators of educational quality: course-taking opportunities, school climate, teacher qualifications, and cognitive stimulation in the classroom. Our findings, reasonably consistent across states, supported both the “home effects” and the “schooling effects” explanations: the hypothesized explanatory variables related to student outcomes as expected. This exercise may be criticized as merely recapitulating

[1] The research reported here was funded by the National Assessment of Educational Progress Data Reporting Program of the National Center for Education Statistics (NCES) under a grant to Michigan State University. The views expressed herein do not represent the position of NCES. This paper summarizes and discusses findings from two papers: “Synthesizing Results from the Trial State Assessment,” to appear in the *Journal of Educational Statistics;* and “Inequality of Access to Educational Opportunity: A National Report Card for Eighth Grade Math,” *Educational Evaluation and Policy Analysis,* 20(4), 253–268. Authors of both papers are Raudenbush, S.W., Fotiu, R.P., and Cheong, Y.F. I wish also to thank Marcy Wallace for administering the many tasks associated with the analysis and Zora M. Ziazi for her work on data analysis.

decades of educational research, and not even with the best available data.[2] Yet TSA does offer the opportunity to compare results across states, for it is the only data set that contains a large, representative sample of students in each of many states.

Perhaps more importantly, the analyses within states bears directly on controversies surrounding accountability at the state level. Our key finding was that, while states vary substantially in unadjusted proficiency means, once we control for NAEP indicators of student background and educational quality, nearly all of the state variation vanishes. This makes sense, in that state-level policies (e.g., regulations, incentives, and aid) can presumably affect student outcomes only by affecting specific educational resources and practices at a more local level, i.e., within schools and classrooms. If those local resources and practices were fully controlled in our models, there would be no direct role for state policy to affect student achievement.

Yet once we verify that state differences almost entirely reflect variation in measurable aspects of student background and school quality, our focus logically shifts to these “correlates of proficiency.” In particular, state differences in correlates of proficiency *that can be manipulated by policy* become especially salient. This led to our second investigation: a study of state-to-state variation in the provision of key educational resources, in particular those resources found consistently related to student outcomes across states.

We were especially interested in equality of access to those resources as a function of student social and ethnic background. Our logic was as follows: having found what many prior studies have found, i.e., that socially disadvantaged and ethnic minority students are at high risk of poor performance, we are inclined to ask about the extent to which these students have access to key resources for learning.

Our results were again not surprising, but nonetheless disconcerting: socially disadvantaged students and ethnic minority students (particularly African American, Hispanic American, and Native American students) are significantly less likely than other students to have access to favorable course-taking oppor-

[2] The cross-sectional data of the TSA do not enable the degree of control for prior student achievement that is possible in a longitudinal study such as NELS. Moreover, NAEP indicators of educational policy and practice are not nearly as refined as are those in NELS.

tunities, school climates, qualified teachers, and cognitively stimulating classrooms. However, what is new and perhaps unique is a second finding based on TSA: the *degree* of social and ethnic inequality of access to resources varies substantially by state. This finding led us to propose a novel “report card” for states based not on mean outcomes, but rather on the extent to which the schools in a state provide key resources for learning. Moreover, our report card allows examination not only of state differences in overall access to these resources, but also state differences in the extent to which access is equitable as a function of social background and ethnicity.

These analyses, while fruitful in our view, also reveal important limitations in data provided by the TSA. These limitations are not so much on the outcome side, where most attention has focused on the construction of NAEP, but rather on the input side. Indicators of student background and especially of key educational resources are currently quite limited in the TSA. For example, student socioeconomic status is indicated by parental education in our analyses. Indicators of parental occupation, income, eligibility for free lunch, and census-based indicators of neighborhood demographic condition, housing, etc., are absent. Regarding school-level organization, NAEP includes indicators of disciplinary climate, but no indicators of staff cohesion, control, and expectations, or of academic press. Indicators of cognitive stimulation in the classroom are few and do not constitute a meaningful or reliable scale. Hence, we settled on a single indicator: emphasis on reasoning during math instruction.

Given the limitations of NAEP indicators of student background, school organization, and instruction, our finding that NAEP indicators can account for nearly all the variation between states was a pleasant surprise. A more refined set of indicators would, however, provide more useful information to those who wish to use TSA, not just to “take the temperature” of the states, but to identify specific targets and strategies for interventions aimed at reducing inequality and thereby improving overall levels of student proficiency.

In the following pages, I aim first to sketch briefly the longstanding debate over sources of educational inequality and its implications for accountability at the state level. Second, I describe the first phase of our investigation: the modeling of student proficiency within states as a function of student background and educational resources. Third, I report results of our second investigation, which focuses on student access to key educational resources in the participating states. A sub-theme in the description of each phase involves challenges of analysis and measurement that also have important implications for future summaries and uses of data from the TSA.

Home and School Differences As Sources of State Inequality in Mathematics Proficiency

The debate about how to interpret the results from the TSA mirrors the longstanding debate about home and school sources of inequality in student outcomes. Social and ethnic inequality in achievement constitutes a troublesome and enduring aspect of schooling in the U.S. Large achievement gaps between students of high and low socioeconomic status (SES) and between European American students, on the one hand, and African American and/or Hispanic students, on the other, have been verified in every major national study of secondary students, beginning with Coleman et al. (1966). Yet researchers have offered contrasting explanations for such inequality.

Home Environmental Inequality

From one standpoint, the school is an essentially neutral learning environment passively allowing sharp inequality in home circumstances to translate into similar inequalities in learning outcomes. Families have long been known to vary substantially in their capacities to provide educational environments that foster school readiness and reading literacy (Fraser 1959; Wolf 1968). Such differences are linked to social status indicators, including income, parental occupation, and parental education (Coleman et al. 1966; Peaker 1967). Parents of high social status are more likely than parents of low social status to have the resources and skills needed to support their children’s academic learning.

If this explanation were completely sufficient to understand observed achievement gaps, variation in student achievement between schools would simply reflect the varied home environments of students attending those schools. Policy interventions aimed at increasing equity might focus primarily on early interventions such as Head Start and on providing support for the families of the most disadvantaged children. Interventions at the classroom or school levels, though perhaps laudable for increasing mean achievement, would hold less promise for reducing inequality.

School Environmental Inequality

From an entirely different standpoint, schools are a much more active force, subjecting essentially similar children to dramatically different learning experiences and thereby actively recreating in each new generation a wide intellectual inequality that conforms to the wide inequalities in earnings and occupational prestige. Clear expositions of this view appear in Ryan (1971), Bowles and Gintis (1976), and Kozol (1991). Tracking (Oakes 1985, 1990), differential teacher expectations (Rosenthal and Jacobson 1968; Rist 1970), and varied school ethos or climate (Rutter et al. 1979), course requirements (Lee and Bryk 1989), teacher subject matter and pedagogical knowledge (Finley 1984; Rosenbaum 1976), and level of cognitive stimulation in the classroom (Page 1990; Rowan, Raudenbush, and Cheong 1993) are aspects of the schooling system often viewed as fostering unequal opportunity and outcomes.

If inequality of schooling were the sole determinant of inequality of educational outcomes, inequality in school mean achievement would reflect school differences in policy and practice. Not surprisingly, those who have emphasized the school as a causal agent in creating educational inequality, while often endorsing compensatory educational policies, have called for sweeping structural reforms in the provision of schooling. These include the elimination of tracking, school finance reform that would equalize spending across rich and poor districts (Berne 1994), and a recasting of teacher preparation to foster more favorable expectations and more cognitively stimulating instruction for currently disadvantaged students. If the "school effects" explanation were correct, such reforms would reduce or eliminate differences between schools in achievement.

The debate reviewed above leaves school differences in student mean outcomes open to vastly different interpretations. One observer might view an elevated school mean as simply reflecting an advantaged school composition; another would attribute this success to excellent school governance, organization, policy, and instructional practice. Those who study school effects seek to measure key aspects of both student composition and school process to assess the relative contributions of each and to isolate those contributors to achievement that reformers can modify (Fuller 1987; Lee and Bryk 1993). Causal inference in such studies is always perilous because student composition and school process are inevitably correlated. Thus, if either student composition or school process is not measured well and is still included in the analysis, estimates of both will be biased.

Given the difficulty of conducting sound studies of school effects, it is not surprising that schemes designed to hold schools accountable for their mean achievement levels have encountered intense criticism (Willms 1992). School means that are not adjusted for student composition will typically convey an overly negative picture of school process in those schools with the most disadvantaged students. However, incorporating adjustments for composition typically leads to underestimates of the effectiveness of schools having favorable student composition (Raudenbush and Willms 1995).[3]

Implications of the Debate for Interpreting State Variation in Outcomes

All of the difficulties in interpreting school differences in mean outcomes are amplified when interest focuses on state mean differences. First, state means are simply aggregates of school means—the same means that have been found difficult to interpret in all but the most careful studies. Second, while all of the problems associated with interpreting either unadjusted or adjusted school means are present in adjusting state means, others are added. For example, the association between student composition and school processes will vary from state to state, as we show below, making the problem of finding meaningful adjustments for student composition even more perplexing. And differences in state means will at least partially reflect differences in state policy. Such policy differences may also be correlated with school composition and school process, creating extra uncertainty about the sources of state variation.

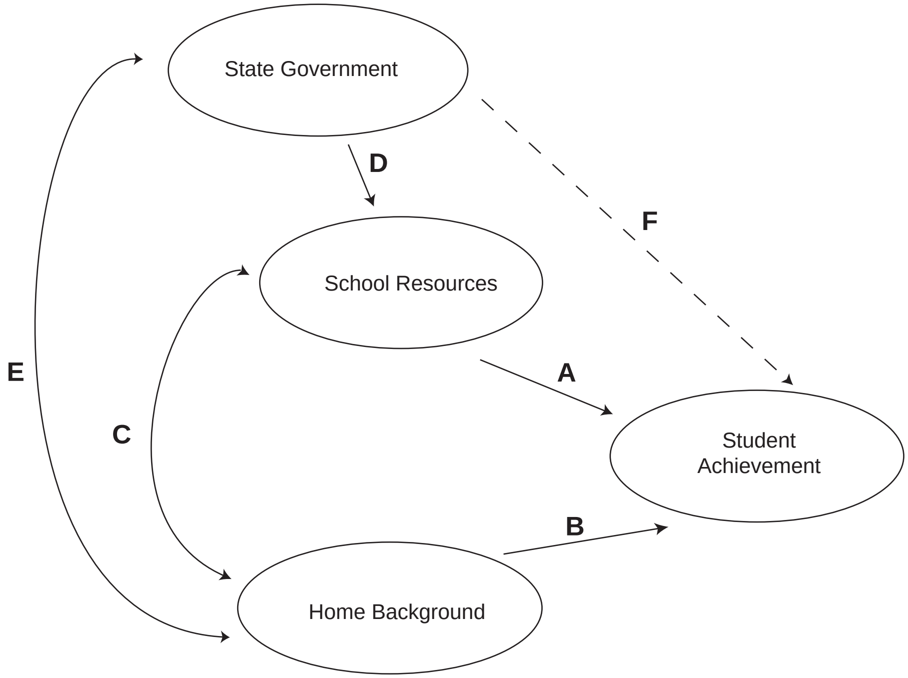
Thus, while making good estimates of state mean proficiency appears essential to any picture of the condition of the nation’s education system, state differences in mean proficiency are, by themselves, intrinsically ambiguous at best and misleading at worst because of the inevitable temptation to make groundless causal inferences.

[3] Student advantage is typically positively correlated with effective school process. Analyses that control student demographics without incorporating good measures of school process will over-estimate the importance of student background, thus leading to overly severe adjustments for student background and thereby underestimating the effectiveness of schools serving advantaged students. Rarely do school accountability studies measure key aspects of school process.

The problem of interpreting state means can perhaps be clarified with reference to a simple causal model (figure 1). Those who interpret state means from TSA are typically interested in the role of state government in improving student achievement (arrow F of figure 1). However, in principle, states cannot directly alter student learning (which is why arrow F is a “dashed line” rather than a solid line). Instead, state policy may affect student achievement *indirectly* by encouraging favorable practice and resources at the level of the school or teacher (arrow D). Schools and teachers can directly affect student achievement (arrow A), though any analysis of such effects must account for student background (arrow B) because school and teacher practice are likely correlated with student background (arrow C).

The first phase of our analysis uses NAEP data to study arrows A and B, i.e., to assess contributing school and teacher quality and the contribution of student background in each of 41 states. The second phase considers arrow D,

Figure 1. Conceptual Model for State-level Policy Effect on Student Achievement



the differences between states with respect to those school and teacher resources and practices found consistently correlated with student achievement.

Phase I: Correlates of Proficiency within States

The first phase of our analysis was to study home and school correlates of eighth grade mathematics proficiency within each state. Our hypotheses were that student social, ethnic, and linguistic background, along with indicators of the home literacy environment, would be related to mathematics proficiency, as in past research; and that indicators of key aspects of school quality, such as course-taking opportunities, disciplinary climate, teacher qualifications, and cognitive stimulation in the classroom, would also predict proficiency. It was essential in this analysis that effects of student background and school quality indicators be adjusted for each other and for other contextual variables such as the composition of the school. This exercise could be viewed as much as a validation study of TSA indicators as a test of theory. We wanted to see whether TSA indicators of home background and school quality were sufficiently well measured to reproduce essential findings of past research. We also sought to examine the power of our within-state models to account for variation between states.

Our expectation was that key variables measured at the student and school level would account for most of the variation between states. This expectation was driven by substantive, rather than statistical, concerns. Controlling for explanatory variables at lower levels of aggregation, such as the student or the school, need not reduce variation at a higher level, such as the state. The adjusted between-state variation can, in principle, be either smaller or larger than the unadjusted between-state variation. However, it stands to reason that states will vary in outcomes for two reasons: selection processes and effects of state educational policy and practice. Selection processes arise because patterns of settlement, fertility, and economic dislocation produce state variation in the demographic and cultural backgrounds of students and their families. Educational policies and practices of schools vary because of the uniquely decentralized character of the U.S. education system and because states and localities tailor the provision of education to the populations they serve. However, states are limited in the “levers” available to them to affect student outcomes. These levers include regulations, incentives, and forms of aid that can have only indirect effects on students by affecting district and school leadership and, ultimately, instruction. It follows that if key aspects of selection, school practice, and instruction are controlled, no state variation will remain to be explained. In terms of figure 1, once arrows A and B are controlled, arrow F should be nonsignificant. This makes sense theoretically but may be difficult to show empirically with NAEP data because NAEP indicators of school resources and home background are limited.

Sample and Measures

Sample

The analyses are based on data from 99,980 eighth graders attending 3,537 schools located in the 41 states and territories participating in the 1992 Trial State Assessment in mathematics. Thus, the average state sample included 2,377 students and 86 schools.

Students within each state were selected by means of a two-stage cluster sample with stratification at the first stage. Specifically, schools were first stratified on the basis of urbanicity, minority concentration, size, and area income; then (a) schools were selected at random within strata with a probability proportional to student grade level enrollment; and (b) students were systematically selected from a list of students, given a random starting point, within schools. It is essential that the analysis plan take into account the stratified and clustered nature of the sample.

Measures

Table 1 lists the variables used and their descriptive statistics. The variables include student outcome data, demographic indicators, home environmental indicators, and classroom and school characteristics.

Measures of math proficiency. The math proficiency data collected as part of NAEP involve a matrix-sampling scheme in which each student was observed on only a subset of relevant items. Rather than yielding a single measured variable, NAEP produces five “plausible values”—random draws from the estimated posterior distribution of each student’s “true” outcome given the subset of items and other data observed on that student (Johnson, Mazzeo, and Kline 1993).

Measures of student demographics. Student demographic variables consist of gender (indicator for male), ethnicity (indicators for Hispanic American, non-Hispanic black American, Asian American, and Native American, with

Table 1. Descriptive Statistics for Student- and School-level Variables for the Combined Sample

Variables	Code and range	Mean	Standard deviation
Student-level data (99,980 students)			
Outcome variables			
Math proficiency 1	(-2.96, 3.06)	0.03	0.99
Math proficiency 2	(-3.82, 2.71)	0.03	0.99
Math proficiency 3	(-3.75, 3.33)	0.03	0.99
Math proficiency 4	(-3.22, 2.87)	0.03	0.99
Math proficiency 5	(-3.84, 2.76)	0.03	0.99
Demographics			
Male	0 = No, 1 = Yes	0.50	0.51
African American	0 = No, 1 = Yes	0.15	0.36
Hispanic American	0 = No, 1 = Yes	0.14	0.35
Asian American	0 = No, 1 = Yes	0.03	0.19
Native American	0 = No, 1 = Yes	0.02	0.12
Not born in U.S.	0 = No, 1 = Yes	0.07	0.26
Student-level data (99,980 students)			
Home environment			
Living with both parents	0 = No, 1 = Yes	0.70	0.47
Living with one parent	0 = No, 1 = Yes	0.20	0.41
Parental education—high school diploma	0 = No, 1 = Yes	0.30	0.47
Parental education—more than high school diploma	0 = No, 1 = Yes	0.18	0.40
Parental education—bachelor’s degree or more	0 = No, 1 = Yes	0.26	0.45
Hours watching TV	(0, 6)	3.17	1.61
Changed school in past 2 years	0 = No, 1 = Yes	0.22	0.42
Get newspaper regularly	0 = No, 1 = Yes	0.73	0.46
More than 25 books in home	0 = No, 1 = Yes	0.91	0.29
Get magazines regularly	0 = No, 1 = Yes	0.76	0.44
Classroom characteristics			
Taking algebra	0 = No, 1 = Yes	0.19	0.40
Taking pre-algebra	0 = No, 1 = Yes	0.25	0.44
Teaching experience of math teacher	(1, 30)	13.44	8.85
Math teacher majored in math	0 = No, 1 = Yes	0.43	0.51

Table 1. Descriptive Statistics for Student- and School-level Variables for the Combined Sample (continued)

Variables	Code and range	Mean	Standard deviation
Math teacher majored in math education	0 = No, 1 = Yes	0.18	0.39
Math teacher did graduate work	0 = No, 1 = Yes	0.47	0.51
Math teacher emphasized reasoning/ analysis in class	0 = otherwise 1 = heavy/moderate	0.46	0.51
School-level data (3,537 schools)			
School-level variables			
Median income (in thousands)	(9.073, 85.567)	28.80	10.73
Instructional dollars per pupil	(7.5, 17.5)	67.22	30.23
Percent minority	(1,100)	28.02	27.70
Urban location	0 = No, 1 = Yes	0.23	0.42
Rural location	0 = No, 1 = Yes	0.23	0.42
Offering 8th grade algebra for high school credits	0 = No, 1 = Yes	0.75	0.43
Availability of computer	0 = No, 1 = Yes	0.83	0.37
School climate	(-3.003, 1.191)	0.00	0.63

European American as the reference group), national origin (indicator for born outside the U.S.), family type (indicators for living at home with a single parent, living at home with both parents, with other type as the reference group), and parental education (indicators for high school graduate, some education after high school, and college graduate, with not graduated from high school or the eighth grader not knowing parents' educational level as the reference group).

Table 1 presents the descriptive statistics on student demographics for the combined 41 states. As table 1 shows, half of the 99,980 students were male. African Americans made up 15 percent of the sample; Hispanic Americans, 14 percent; Asians, 3 percent; and Native Americans, 2 percent; and 7 percent of the students were not born in the U.S. In addition, 70 percent of the students indicated that they had two parents residing at home, and 20 percent of students reported that they lived in a single-parent household. For 30 percent of the sample, either the mom or the dad held a high school diploma; for 18 percent, one parent had some education after high school graduation; and for 26 percent, at least one parent graduated from college.

Measures of home environment. Home environment variables include amount of time watching television, mobility (as indexed by whether a student changed schools in the past two years), home literacy environment (indicators for receiving a newspaper, having more than 25 books, and subscription of magazines). Table 1 indicates that the students spent 3.17 hours daily on average watching TV. Less than a quarter of them (22 percent) reported that they had changed schools in the past two years. About three-fourths of the students (73 percent and 76 percent) indicated that their households regularly got newspaper and magazines, respectively. The great majority of the students, 91 percent, had more than 25 books in their homes.

Measures of classroom characteristics. Classroom characteristics involve type of course (indicators for pre-algebra, algebra, with other course as the reference group), the teaching experience and qualifications of the teacher of the student (indicators for undergraduate math major in college, math education major in college, with other major as the reference group; and an indicator for having a graduate degree), as well as teacher-reported emphasis on reasoning in the classroom (an indicator for moderate to high emphasis). The data on teacher background and pedagogical practice were taken from responses to questionnaires administered to the mathematics teachers of the students sampled.

Table 1 shows that 19 percent of the students in the sample enrolled in an algebra course and 25 percent of them took pre-algebra. The average number of years of teaching experience for the teachers of the students sampled was about 13. Furthermore, 43 percent of the students had a teacher who majored in mathematics as an undergraduate; 18 percent of the students had a teacher who was a math education major; and 47 percent of the students had a teacher who got a graduate degree. About half of the students (47 percent) attended a classroom where reasoning received moderate to high level of emphasis.

Measures of school characteristics. School characteristics include the social and racial composition of a school as measured by median income and percent minority (Hispanic and African American students). Other school-level measures are location (indicators for an urban school, a rural school, with suburban school as a reference group), and financial and computing resources as indexed by instructional dollars per pupil and availability of computers (an indicator for the availability of computers in a math classroom or a lab for most of the time), course offerings (an indicator for the availability of algebra for high school credit), and a scale measuring the disciplinary climate of the school. The scale was created from the following items indicating the extent to which each was a problem in the school: tardiness, absenteeism, cutting classes, physical conflicts, drug and alcohol use, health, teacher absenteeism, racial or cultural conflict. Each item was first standardized, and the scale was constructed as the average of the nine standardized scores. Average Cronbach's alpha for the 41 states was .79.

Analytic Approach

Math Proficiency

Our strategy for modeling math proficiency has two stages: a within-state analysis and a between-state analysis. The within-state analysis uses a hierarchical linear model to handle the clustered character of the sample. Sample design weights are applied at the student level to accommodate the stratified character of the sample and the associated over-sampling of certain subgroups. This analysis is replicated for each plausible value and the results pooled as recommended in Little and Schenker (1994) and Mislevy (1992), using a specialized version of the HLM program (Bryk, Raudenbush, and Congdon 1994) originally adapted for multiple plausible values by Arnold, Kaufman, and Sedlacek (1992). The output for each state is a vector of parameter estimates and their estimated sampling variance matrix. These then provide input data for the second stage of the analysis, which involves an empirical Bayes and a Bayesian synthesis of findings across states. The syntheses employ the method of moments (Raudenbush 1994) and the Gibbs sampling (Gelfand and Smith 1990). (See Raudenbush, Fotiu, and Cheong [1998] for a full exposition of the approach.) Taken together, the two stages have the structure of a planned "meta-analysis" (Glass 1976) in which each state's separate analysis constitutes a "study," and the between-state analysis combines these results.

Within-state Models

To address these questions, we first formulated within each state two separate two-level hierarchical models, one with and one without covariates (measures on student demographics, home environment, and classroom and school characteristics). Past research on the associations between the social distribution of educational resources and outcomes guided the specification of the former model (e.g., Bernstein 1970; Bryk and Thum 1989; Coleman et al.1966; Finley 1984; Oakes 1985; Page 1990; Raudenbush, Rowan, and Cheong 1993; Rosenbaum 1976; and Rutter et al. 1979). The model is

1. $$Y_{ijk} = \beta_{0k} + \sum_{p=1}^{p} \beta_{pk} X_{pijk} + u_{jk} + e_{ijk} \, ,$$

where

Y_{ijk} is the math proficiency score for student *i* in school *j* and state *k*;
β_{0k} is the mean for state *k*, which is adjusted for the school- and student-level covariates;
X_{pijk} is the p^{th} covariate, which is centered around the Michigan mean;
β_{pk} is the regression coefficient associated with each X_{pijk};
u_{jk} and e_{ijk} are the residual random school and student effects. They are assumed independently and normally distributed with ω_k^2 and σ_k^2 respectively.

Estimates of the two variance components, ω_k^2 and σ_k^2, incorporate variation associated with the cluster sample so that the maximum likelihood (ML) estimate of each regression coefficient and its standard error incorporates the extra variation arising from the clustered nature of the sample. The use of sampling weights accounts for unequal probability of selection and multiple plausible value analysis accounts for the estimation of proficiency.

Deviating the school- and student-level covariates around the Michigan means allows us to obtain more precise estimates of various parameters for our own state, Michigan.[4] For the sake of simplicity, we forego the option of allowing any of the partial effects associated with student-level covariates to vary randomly from school to school within state *k*. Thus, only β_{0k}, the intercept, varies randomly across schools within states.

Between-state Models

The between-state synthesis combined the output produced by each state to obtain inferences on parameters for individual states as well as global parameters. The output from the within-state analysis for state *k* consisted of the ML estimates b_k of the state mean and its estimated sampling variance v_k. The estimate b_k is assumed to vary around its corresponding parameter β_k with an

[4] The covariates can be deviated around other constants such as the national means for other purposes.

unique error r_k associated with the sample for state k, which has a known sampling variance v_k, i.e.,

2. $b_k = \beta_k + r_k , r_k - N(0,v_k).$

The parameter β_k is in turn assumed to vary around an overall mean γ plus a random error associated with state k, μ_k. We may write

3. $\beta_k = \gamma + \mu_k , \mu_k - N(0,\tau_k).$

The random error has a variance of τ.

Table 2 lists the approximate posterior means and standard deviations of the various regression coefficients, and the estimates of between-state variance and their square roots.[5] We computed z-ratios for the regression coefficients to evaluate the null hypothesis that a particular regression coefficient pooled across states was 0. A z-ratio larger than 2 or 3, as indicated by asterisks in table 2, lent support to rejection of the null hypothesis.

Student demographics. Controlling for home environments and for classroom and school characteristics, the results suggest that, on average, males had higher scores than females; and African Americans, Hispanic Americans, and Native Americans exhibited lower proficiency than did European or Asian Americans. For instance, African Americans obtained, on average, about half a standard deviation lower math proficiency than did European Americans. Net of other covariates, students who were born in the United States scored higher that those who were not. The partial effects associated with the African American and Hispanic American ethnicity and the place of birth variables seem to vary from state to state.

Home environment. Controlling all other covariates, family structure, parental education, and home literacy environment were related to proficiency. Students who lived with either one parent or both parents outperformed those who did not and also those who did not know the educational levels of their parents. Students whose parents had education beyond high school and those whose parents had college degrees scored higher than did those whose parents had not graduated from high school. Furthermore, students coming from house-

[5] Table 2 gives the empirical Bayes summary results. Raudenbush et al. (in press) provide results from the fully Bayesian synthesis and compared the two sets of results. Individual state results are available upon request.

Table 2. Empirical Bayes Summary of State-by-State Results

Predictors	Approximate posterior mean of γ_p	Approximate posterior standard deviation of γ_p	Estimate of between-state variance, τ_p	Square root of the estimate of between-state variance, $\tau_p^{1/2}$
Demographics				
Male	0.0904*	0.0061	0.0005	0.0231
African American	-0.4583*	0.0215	0.0123	0.1107
Hispanic American	-0.3894*	0.0271	0.0239	0.1546
Asian American	0.1288*	0.0216	0.0032	0.0565
Native American	-0.2162*	0.0223	0.0043	0.0654
Not born in U.S.	-0.2369*	0.0211	0.0110	0.1046
Home environment				
Living with both parents	0.2884*	0.0116	0.0021	0.0456
Living with one parent	0.2500*	0.0124	0.0022	0.0471
Parental education—high school diploma	0.0567*	0.0082	0.0008	0.0273
Parental education—more than high school diploma	0.2455*	0.0085	0.0217	0.2455
Parental education—college degree	0.2146*	0.0125	0.0041	0.0638
Hours watching TV	-0.0404*	0.0024	0.0001	0.0112
Changed school in past 2 years	-0.0640*	0.0063	0.0000	0.0000
Get newspaper regularly	0.0277*	0.0057	0.0000	0.0000
More than 25 books in home	0.2051*	0.0092	0.0000	0.0000
Get magazines regularly	0.1006*	0.0075	0.0007	0.0259
Classroom characteristics				
Taking algebra	0.9830*	0.0201	0.0141	0.1188
Taking pre-algebra	0.3972*	0.0159	0.0083	0.0912
Teaching experience of math teacher	0.0029*	0.0006	0.0000	0.0000
Math teacher majored in math	0.0844*	0.0121	0.0038	0.0844

Table 2. Empirical Bayes Summary of State-by-State Results (continued)

Predictors	Approximate posterior mean of γ_p	Approximate posterior standard deviation of γ_p	Estimate of between-state variance, τ_p	Square root of the estimate of between-state variance, $\tau_p^{1/2}$
Math teacher majored in math education	0.0823*	0.0149	0.0055	0.0738
Math teacher did graduate work	0.0101	0.0084	0.0010	0.0320
Math teacher emphasized reasoning/analysis in class	0.1373*	0.0096	0.0023	0.0478
School characteristics				
Median income	0.0059*	0.0007	0.0000	0.0000
Instructional dollars per pupil	0.0000	0.0000	0.0000	0.0000
Percent minority	-0.0036*	0.0000	0.0000	0.0000
Urban location	0.0140	0.0143	0.0014	0.0380
Rural location	-0.0191	0.0225	0.0125	0.1120
Offering 8th grade algebra for high school credits	-0.0425*	0.0138	0.0018	0.0428
Availability of computer	0.0024	0.0124	0.0000	0.0000
School climate	0.0378*	0.0079	0.0000	0.0000
Intercept				
Intercept	0.0680	0.0096	0.0000	0.0000

*** z-score > 3.**

holds that had more than 25 books in the home and received newspaper and magazines regularly had higher math proficiency than those who came from households that did not. There were statistically significant negative partial effects associated with time spent watching TV and changing school in the past two years. Three of the between-state variance estimates were 0.

Classroom characteristics. Enrollment in algebra and pre-algebra were positively related to math scores, all else being equal. Those who took algebra scored about one standard deviation higher than the reference group, whose students took eighth grade math or other non-algebra course or who did not take any math course. Those who enrolled in pre-algebra scored about 0.4 standard deviation higher than the reference group. Teaching experience, teacher subject matter expertise (as indicated, respectively, by majoring in math or math education), and emphasis on reasoning[6] were also positively correlated with proficiency in math, net of the effects of other covariates.

School characteristics. School composition effects were manifest, net all other predictors, including student demographic background. In particular, school median income was positively related to proficiency, and percent minority was negatively related to proficiency. Thus, school social class and ethnic segregation effects tend to reinforce differences based on individual social class and ethnicity. All else being equal, a favorable school climate was positively related to proficiency. The estimated partial effect of school algebra was statistically significant and negative. Note that this effect represented the expected difference in math proficiency between a student not taking algebra in a school that offered algebra and a student in a school that did not offer algebra. One implication of the predominantly negative effect across the states is that there are at least some students in schools not offering algebra who would have benefited from enrollment in an algebra course had they attended schools that did offer algebra. In addition, as taking algebra was, in general, the most powerful single predictor of proficiency, one must conclude that attending a school that offers algebra is related positively to math proficiency.

In sum, the relevant covariates include indicators of student demographic status, home environment, and school composition; these relate to proficiency as expected. At the school level, a curriculum that includes opportunities to take high school algebra and a positive climate were linked to proficiency. At the classroom level, teachers' subject-matter preparation, as indicated by hav-

[6] One would expect the level of reasoning to increase with teacher's education (e.g., a teacher's undergraduate major) and the difficulty of the course (e.g., an algebra course versus a general mathematics course). Emphasis on reasoning, teacher's education, and course type thus may jointly influence math proficiency. To understand how these various predictors may be correlated with the math scores, two models were specified, one with and one without emphasis on reasoning entered as a predictor. The results showed that reasoning, independent of all other covariates, was positively related to math proficiency. In fact, the estimates of other predictors remained nearly the same in the two models.

ing majored in math or in math education, and emphasis on mathematical reasoning predicted elevated proficiency.

Variance Reduction

Figures 2 and 3 give the approximate marginal posterior for the variance for the intercept, that is, for var(β_{0k}) = τ for the unconditional (with no covariates) and conditional models.[7] Figure 3 shows unmistakable evidence of heterogeneity between states (note that 0 is not a plausible value for τ). However, there is considerable uncertainty about the magnitude of this heterogeneity.

The math proficiency measure was on a scale with a mean near 0 and a variance of approximately unity. The posterior mean of τ is .088, implying that about 8.8 percent of the variance in the outcome lies between states. However, τ values as small as .04 and as large as .14 are not improbable. Thus, it appears that from 4 percent to 14 percent of the variance in the outcome lies between states.

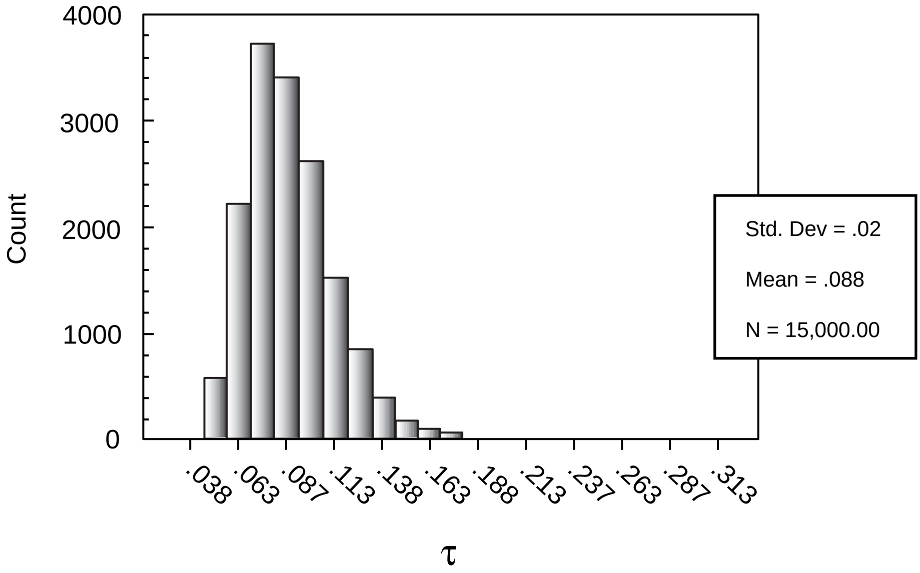
Whereas figure 3 shows evidence of heterogeneity between states (note that 0 is not a plausible value for τ) after controlling for the various measures, there is every reason to believe that the magnitude of this heterogeneity is small. The posterior mean of τ is .018, implying that 1.8 percent of the variance in the outcome lies between the intercepts of the states. Moreover, the unknown value of τ is unlikely to exceed .03 or 3 percent of the total variance in the outcome. It appears that from .004 percent to 3 percent of the variance in the intercept lies between states after controlling for covariates. Thus, most of the state-to-state heterogeneity is explainable on the basis of covariates defined on students, teachers, and schools. This indicates, in general, that states with high mean proficiency tend to be advantaged on the relevant covariates and that these advantages account for most state-to-state variation in proficiency.

Phase II: Inequality of Access to Educational Opportunity

In terms of figure 1, our “first phase” analysis found certain school resources (arrow A) and student background indicators (arrow B) to be quite

[7] The figures are output obtained from the Bayesian synthesis (see Raudenbush, Fotiu, and Cheong [in press] for a description of the approach).

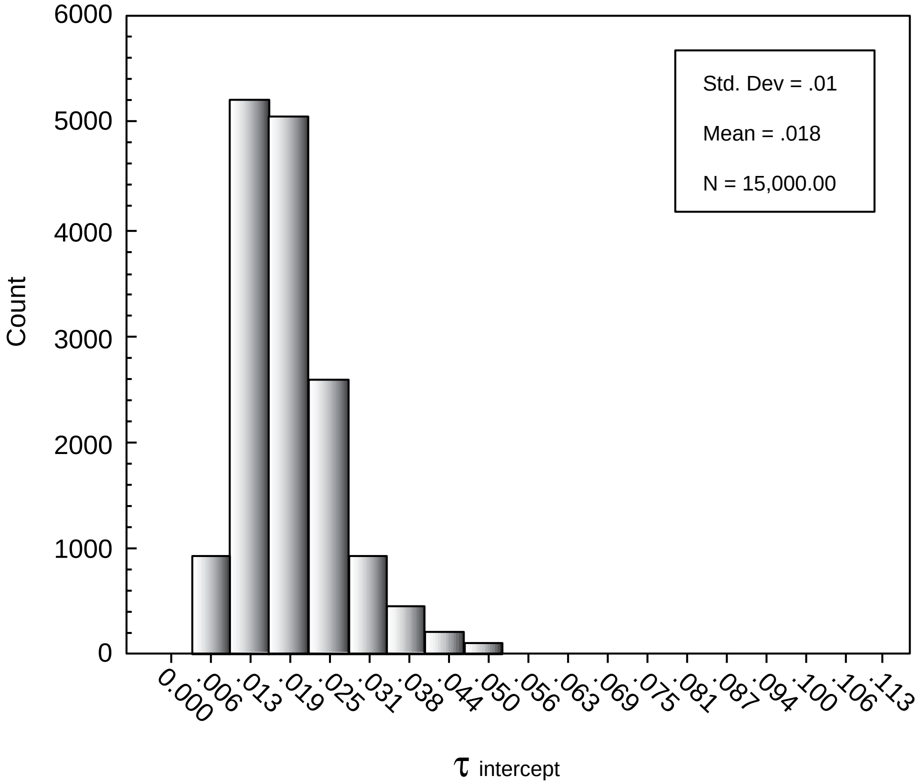
Figure 2. Estimated Posterior Distribution of τ: Unconditional Model



consistently related to student achievement. Controlling for these, state differences in achievement (arrow F) became small, perhaps negligible. This encouraged us to abandon further investigation of state means, whether adjusted or unadjusted. Rather, we sought in Phase II of our investigation to examine state differences in school resources. Given the consistent association between advantaged home background and achievement, we were especially interested in the equity with which the school resources are distributed. We asked: "Does the distribution of school resources likely reinforce or counteract inequalities arising from home environment? Do states differ, not only in the provision of resources, but also in the equity with which they are distributed?"

One product of this work is a different kind of "report card" for states than is typically made available to policymakers. The typical report card provides unadjusted differences between states in academic proficiency. This typical report card, though conveying some useful information, can easily mislead. It tends to provide an overly negative portrayal of education systems in states with comparatively disadvantaged demographics and an overly rosy picture of

Figure 3. Estimated Posterior Distribution of τ intercept: Conditional Model



education in states with more advantaged students. Moreover, it provides little insight into ways in which policy changes might produce better outcomes.

The report card we present compares states on educational opportunities, resources, or processes theoretically and empirically linked to outcomes. It reveals the equity with which these are distributed as a function of student social background and ethnicity. It therefore points the discussion toward interventions that would increase the quality and equity of education provision.

In modeling the relationship between student demographic background and educational resources, our analysis strategy depended on whether the educational resource in question was measured dichotomously or continuously. Dichotomous resources included school course offering (1 = school offers high school algebra, 0 = school does not offer high school algebra) teacher education (1 = teacher majored in math, 0 = teacher did not major in math), and emphasis on reasoning in the classroom (1 = high, 0 = other).

Model for the Continuous Outcome (Disciplinary Climate)

The method of estimation for the model studying school climate involves a two-level hierarchical linear model (Bryk and Raudenbush 1992) with students nested within states. Robust standard errors were computed using the generalized estimating equation approach of Zeger, Liang, and Albert (1988). These standard errors are relatively insensitive to mis-specification of the variances and covariances at the two levels and to the distributional assumptions at each level. State-specific effects were estimated via empirical Bayes (Morris 1983; Raudenbush 1988).

Specifically, we estimated a within-state model in which ethnicity, parental education, and the ethnicity-by-parent interaction predicted school climate. Ethnicity was represented by four dummy variables and parental education by two dummy variables. Allowing for the ethnicity-by-parent interaction effect enabled us to model access to resources for each sub-group (e.g., African Americans of low, middle, or high parental education). We allowed coefficients for the parental education dummies and for African American and Hispanic American ethnicity to vary randomly over states, thus allowing state-by-state comparisons. Sample sizes of Asian Americans and Native Americans were, unfortunately, too small to allow such a fine-grained analysis.

Models for the Dichotomous Resource Indicators

The same explanatory model for the school climate was specified for each dichotomous outcome. In this case, however, we used a two-level logistic regression model, estimated by penalized quasi-likelihood (Breslow and Clayton 1993), with robust standard errors. Such a model is equivalent to a 2 by 3 by 5 by 41 contingency table with 2 levels of the outcome, 3 levels of parent education, 5 levels of ethnicity, and 41 levels representing states.

Results

We now consider the degree of ethnic and social equality in access to the four resources of interest. Specifically, we ask the following questions for each resource indicator:

1. Averaging within the 41 states participating in the TSA, to what extent does student social background, as indicated by parental education and student ethnicity, predict access to the resources?
2. Does the degree of inequality in access vary by state? If so, how do the 41 states compare?

Results Averaged Across States

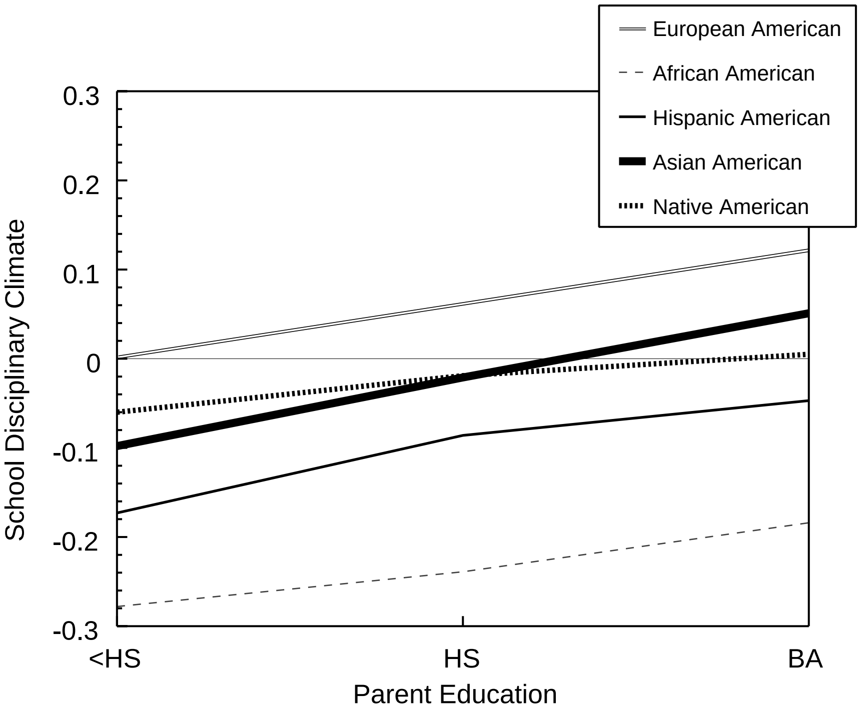
School Disciplinary Climate

Figure 4 gives the graph of the fitted model in which ethnicity and parental education predict access to favorable disciplinary climate. The figure shows that higher levels of parental education are clearly linked to more favorable disciplinary climate. The near parallelism of the five lines (with the exception of the line for Native Americans, which is based on a comparatively small sample) reflects the absence of any statistical evidence of a two-way interaction involving parental education and ethnicity. There is a substantial significant vertical displacement between ethnic groups. Pairwise comparisons using a Bonferroni adjustment to control the family-wise Type I error rate at the 5 percent level indicated four separate clusters of means (in descending order of magnitude): (a) European Americans; (b) Asian Americans and Native Americans; (c) Hispanic Americans; and (d) African Americans. Given that the school climate outcome had a mean of 0 and a standard deviation of 0.63, the differences manifest in figure 4 are non-trivial in magnitude: About 0.20 standard deviation units separate those with parents having a BA from those whose parents were without a high school diploma; nearly half a standard deviation separates European Americans and African Americans.

Access to High School Algebra

Figure 5 plots the predicted probability of attending a school that offers high school algebra for eighth graders as a function of parental education for

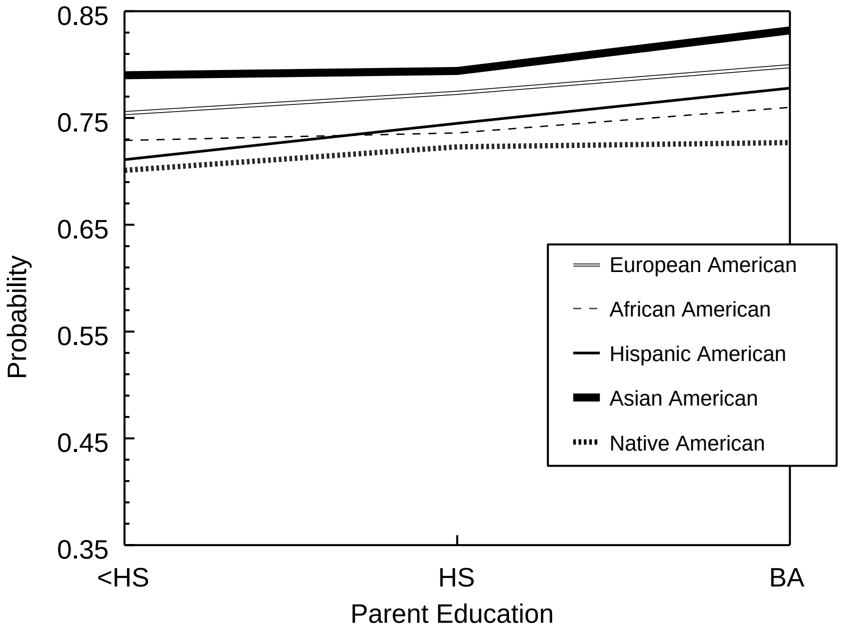
Figure 4. Predicted School Disciplinary Climate as a Function of Parent Education and Ethnicity



each of the five major ethnic groups under study. We see that parental education is positively associated with the probability of attending such a school. As in the case of climate, the near parallelism of the five lines reflects the absence of any statistical evidence of a two-way interaction involving parental education and ethnicity. Again, we find a significant vertical displacement between ethnic groups. Pairwise comparisons using a Bonferroni adjustment to control the family-wise Type I error rate at the 5 percent level indicated three separate clusters of ethnic group probabilities (in descending order of magnitude): (a) Asian Americans; (b) European Americans, African Americans, and Hispanic Americans; and (c) Native Americans. The differences manifest in figure 2 are comparatively modest in magnitude.

The regression coefficients for the predictors give the associated partial effects in terms of log-odds. Besides computing predicted probabilities based on the regression coefficients, one could compute odds ratios as well. For in-

Figure 5. Predicted Probability of Assignment to a School That Offers Algebra as a Function of Parent Education and Ethnicity



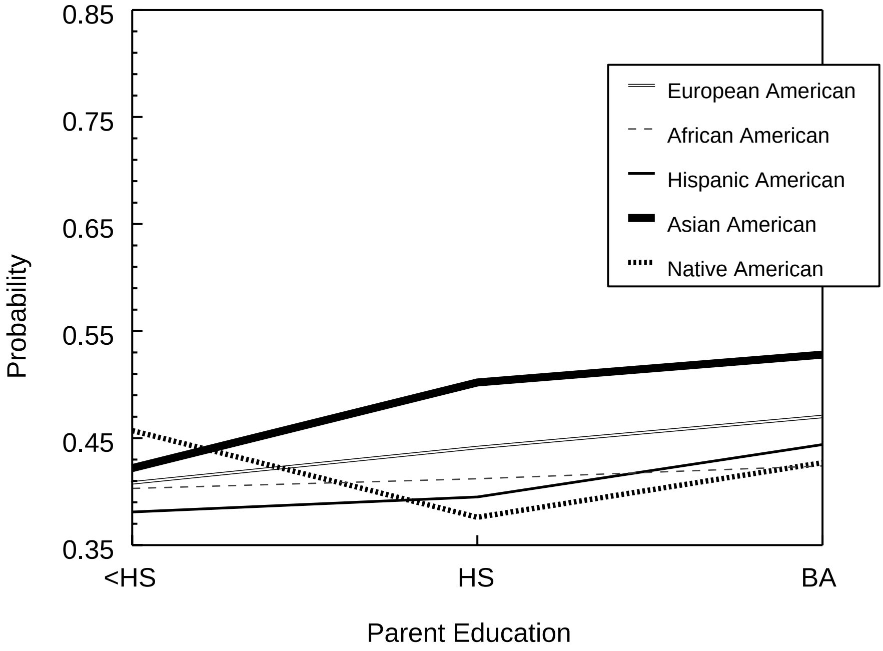
stance, the odds ratio of offering algebra for a school attended by a student whose parent had college education versus a school attended by a student whose parent had less than high school education is $\exp\{\acute{\alpha}_{BA}\} = \exp\{-0.244\} = 0.784$.

We now turn to two classroom-level resources for learning: teacher subject matter preparation, as indicated by having majored in mathematics, and a cognitively stimulating environment, as indicated by an instructional emphasis on mathematical reasoning. In both cases, we find that social background (as indicated by parental education) and ethnicity are linked to access to the resource. However, the findings are more complex than those reported above, in that a two-way interaction is manifest in the case of these two classroom-level resources.

Teacher Preparation

Figure 6 plots the predicted probability of encountering a math teacher who majored in math as a function of social background and ethnicity. The figure shows that higher levels of parental education are linked to a higher

Figure 6. Predicted Probability of Assignment to a Teacher Who Majored in Math As a Function of Parent Education and Ethnicity

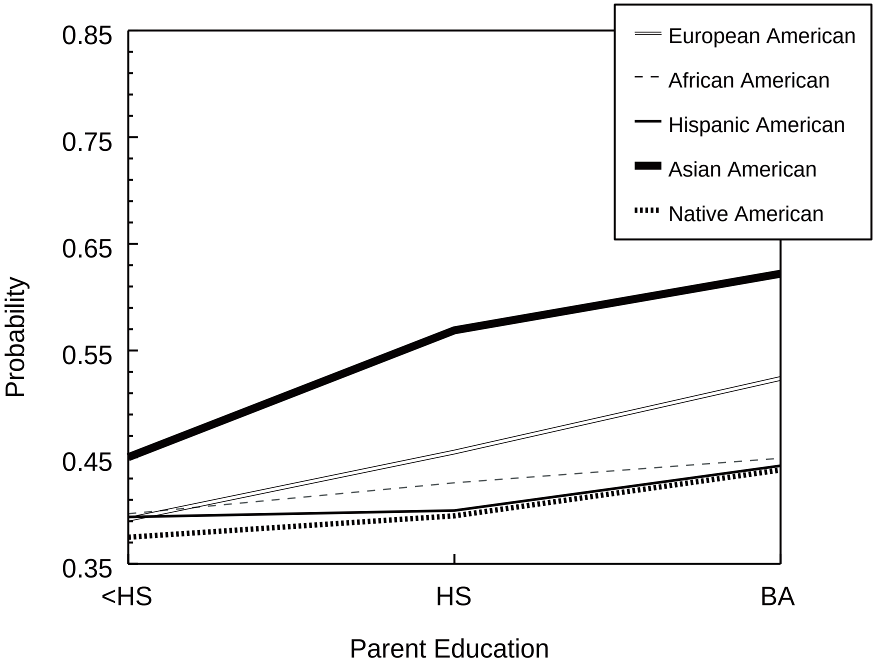


probability of encountering such a teacher. However, the magnitude of this relationship depends upon ethnicity. The link between social background and teacher preparation is strongest for Asian Americans and European Americans and weakest for African Americans, Hispanic Americans, and Native Americans. Equivalently, we can say that ethnic gaps in access to the resource are manifest, but are more pronounced at higher than at lower levels of parent education.

Emphasis on Reasoning

Figure 7 plots the predicted probability of encountering a math teacher who emphasizes mathematical reasoning during instruction. Again there is a positive relationship between parent education and this probability, but again the magnitude of this association depends upon ethnicity. The link between parental education and access to reasoning is strongest for Asian Americans and European Americans and weakest for the other three groups. Equivalently,

Figure 7. Predicted Probability of Assignment to a Math Teacher Who Emphasizes Reasoning As a Function of Parent Education and Ethnicity



just as in the case of teacher preparation, we can say that ethnic gaps in access to the resource are manifest, but are more pronounced at higher than at lower levels of parent education.

Summary

In sum, we find evidence of ethnic and social inequality in access to all four resource indicators when averaging across the 41 states. Main effects of both ethnicity and social background generally parallel previous findings in predicting student achievement. Thus, just as high parental education predicts favorable outcomes, it also predicts access to schools with favorable climates, schools that offer algebra, teachers with training in mathematics, and classrooms that emphasize reasoning. Similarly, ethnic groups disadvantaged in outcomes (African Americans, Hispanic Americans, and Native Americans) also encounter less access to these resources for learning.

State Variation in Access to Resources

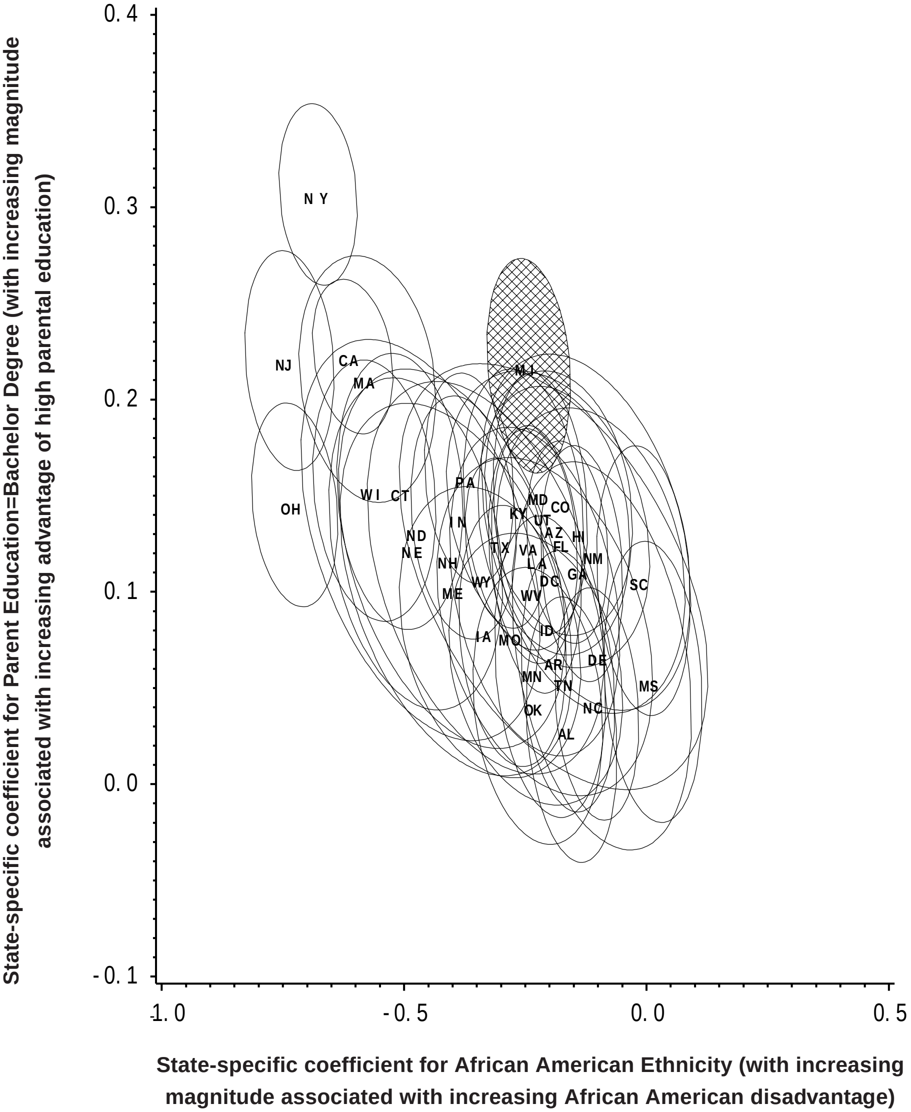
The pooled, within-state findings regarding social and ethnic inequality in access to a favorable school climate provide an “on-average” picture of inequality in access to resources over 41 states. However, these on-average results poorly represent the picture that we find in many states. In fact, the data reveal substantial evidence of state variation.

The case of school disciplinary climate illustrates the substantial variation across states. Figure 8 plots 95 percent bivariate confidence ellipses for the 41 states where the vertical axis is social inequality (as indicated by mean gaps in school climate between students having parental education of BA and less than high school) and the horizontal axis is ethnic inequality (as indicated by mean differences between African Americans and European Americans.[8] Four features of the scatter plot of ellipses are noteworthy:

1. First, there is a rather strong negative relationship between parental education “gaps” and ethnicity “gaps.” That is, states with a high degree of social inequality tend to also exhibit a high degree of ethnic inequality. New York is a case in point; lying in the upper left quadrant, New York has a “parental education gap” of about 0.30 points (half a standard deviation) and an “ethnicity gap” of around 0.60 (a full standard deviation).
2. Some degree of inequality is present in nearly all states. This inference is based on noticing that nearly the entire scatter of ellipses lies above 0 on the vertical axis (indicating positive parental education effects within states) and below 0 on the horizontal axis (indicating that African American ethnicity is associated with lower levels of disciplinary climate).
3. However, the magnitude of inequality varies quite substantially across states. There is a cluster of states near the origin (the point indicating equality on both parental education and ethnicity). There are also states far from the origin (e.g., New York, New Jersey, California, and Massachusetts), implying substantial inequality in access to favorable disciplinary climate in these states.

[8] The mean differences associated with social inequality are adjusted for ethnicity, and the mean differences associated with ethnicity are adjusted for parent education. The 95 percent confidence ellipses are based on the empirical Bayes posterior distribution (Morris 1983) of the parental education and ethnicity coefficients for each state.

Figure 8. 95 Percent Bivariate Confidence Ellipses for the State-specific Coefficients Associated with Parental Education and African American Ethnicity (Outcome: Mean School Climate)



4. There is considerable overlap among the ellipses, making it hard to distinguish many pairs of states and, in fact, making pairwise comparisons confusing. However, the ellipses of any pair of states can be shaded (as Michigan's ellipse in figure 8) to facilitate a desired pairwise comparison. Using computer graphics, it is easy to highlight any subset of states to generate clearer comparisons.

The value of the ellipses is that they automatically communicate the degree of uncertainty about rankings among states. Consider, for example, Michigan and Ohio. Ohio is characterized by significantly greater ethnic inequality than Michigan is, i.e., the gap between European Americans and African Americans in the disciplinary climates they encounter is statistically greater in Ohio than in Michigan, as indicated by the fact that the two ellipses do not overlap on the horizontal axis. However, the two states do not differ in social inequality, as indicated by the fact that their ellipses do overlap on the vertical axis.

Excellence versus Equality

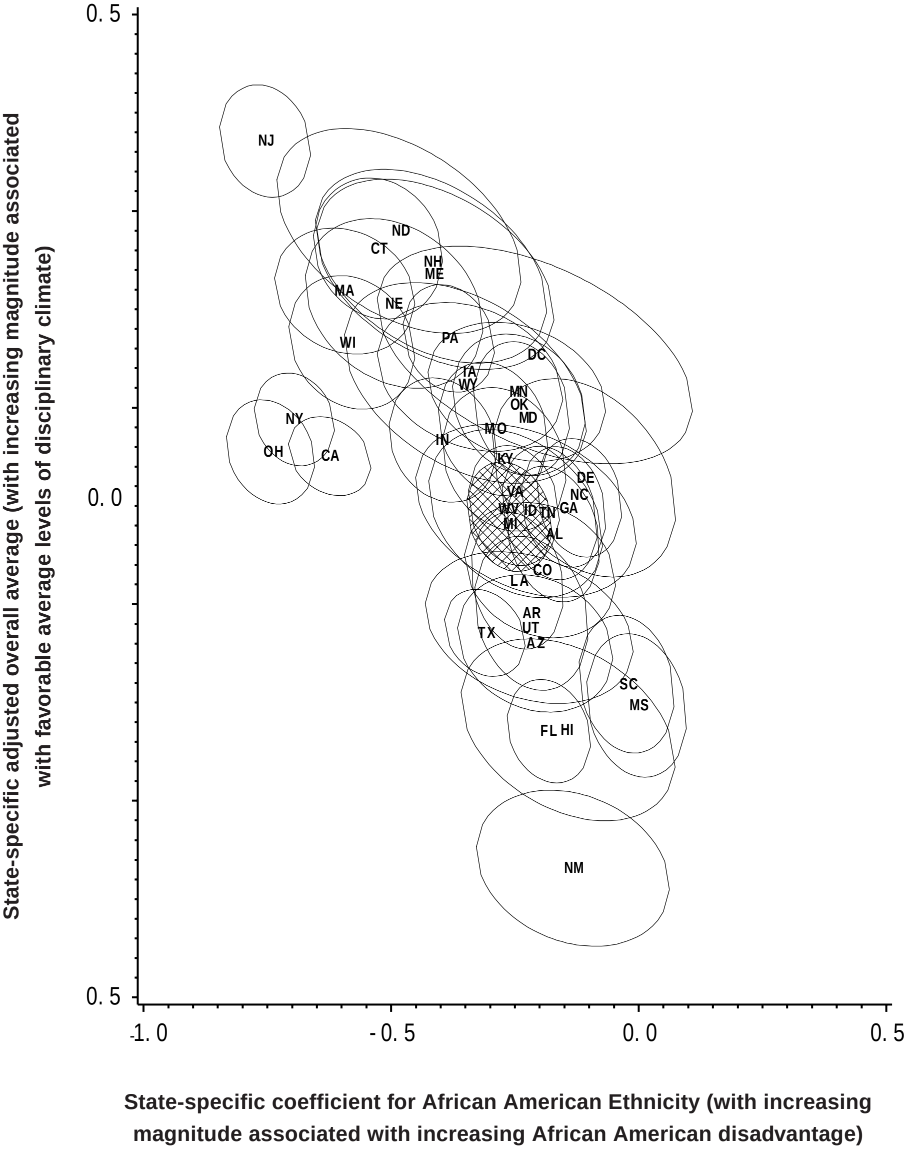
It is also possible to plot "excellence" (high levels of a resource) against "equality," as depicted in figure 9. The figure shows, for example, that New Jersey, though displaying a comparatively high degree of ethnic inequality, has one of the highest average levels of disciplinary climate. Equality is not a good thing if environments are equally bad; South Carolina and Mississippi exhibit low levels of inequality but also low average levels of disciplinary climate.

For the other resources, the pooled results also poorly represent the degree of inequality in some states. Again, the data reveal substantial evidence of state variation. It is possible and generally useful to describe state-to-state variation in access to these resources as we did in the case of school climate (figures 4 and 5). However, a detailed discussion of differences among the 41 states on all resources goes beyond the scope of this paper.

Conclusions

The Trial State Assessment of NAEP reports mean student proficiency in a given subject for each of the participating states, broken down by ethnicity and parental education (c.f., Mullis et al. 1993). Although reports of state means are essential as part of an assessment of the condition of education in the U.S., we have argued in this paper that such state means, by themselves, are difficult

Figure 9. 95 Percent Bivariate Confidence Ellipses for the State-specific Coefficients Associated with Intercept and African American Ethnicity (Outcome: Mean School Climate)



to interpret and even misleading. The means reflect an unknown mix of contributions from student demographics, school organization and process, and state policy. To supplement the reporting of means, we have proposed a reporting of the access that states provide to key resources for learning. Knowing the extent to which states provide these resources to students of varied social background and ethnicity points toward sharply defined policy debates concerning ways to improve education. The results of our analysis are both substantive and methodological.

Substantive Findings

Our results indicate substantial inequality in access to resources, on average, over the 41 participating states. Social background, as indicated by levels of parental education, is significantly related to access to a school with a favorable disciplinary climate and a school that offers high school algebra for eighth graders. Social background also predicts the probability that an eighth grader will encounter a teacher who majored in mathematics and a teacher who emphasizes reasoning during mathematics instruction. These effects of social background are adjusted for ethnicity.

The results for ethnicity parallel those for social background, though they vary to some degree by the resource of interest. For example, with respect to school disciplinary climate, European Americans encounter, on average, the most favorable disciplinary climates; Asian Americans and Native Americans are next, followed by Hispanic Americans and finally by African Americans. The probability of attending a school that offers algebra is distributed a little differently: Asian Americans experience the highest probability of attending such a school; European Americans, African Americans, and Hispanic Americans are next most likely to attend such a school; and Native Americans have the lowest probability of attending such a school. These effects of ethnicity are adjusted for social background. The results for teacher preparation and emphasis on reasoning are more complex: ethnic gaps in access are greatest at highest levels of parental education, with Asian Americans and European Americans having greater access than other groups to each resource.

In sum, we have found substantial evidence of inequality in access to these resources as a function of social background and ethnicity. However, there is also substantial variation across states in the extent of inequality. While some degree of both forms of inequality appears to exist in nearly all states, inequality is much more pronounced in some states than in others. Moreover, the overall level of availability of each resource also varies from state to state. While a fine-grained analysis of state differences on all four resources would be of interest, such a study goes beyond the scope of the current paper. However, we have suggested ways in which state differences might be examined.

The policy implications of these findings vary as a function of the resource in question. Whether a school offers algebra to eighth graders is amenable to direct influence by state and district policy. The key impediment to offering algebra in a given setting is cost. It is generally more costly for smaller schools than for larger schools to diversify their curricula. Similarly, hiring teachers with serious college-level preparation in mathematics is under the direct control of policy, with cost again being a key impediment.

Constructing a favorable disciplinary climate, in contrast, is only partially under the control of policymakers. Effective adult leadership in a school setting is arguably the primary ingredient in creating such a climate, though the active participation of students and parents is also required for success. Skill, knowledge, and commitment are required, and there is considerable uncertainty about how to foster the needed efforts. Similarly, a decision to emphasize reasoning is in the hands of the teacher, depending on the teacher’s knowledge, skills, and evaluation of student needs. Interventions to encourage instruction that emphasizes reasoning are currently widespread, but the outcomes of such interventions are inevitably uncertain.

In sum, how information from a report such as ours ought to influence the policy debate will vary as a function of the kind of resource in question. Options for increasing access to certain resources must be evaluated in terms of cost and feasibility. Our primary point, however, is that systematically collected data on access to key resources, as a supplement to reports of mean proficiency, ought to constitute an important input into policy debates regarding educational reform.

Methodological Implications

The educational resources considered here clearly constitute a small subset of those that ought to be studied. We have reasoned that the resources of key interest are those suggested by prior theory and research and operationalized in NAEP. There should also be some evidence that the NAEP indicator of the resource relates as expected to key educational outcomes. The logic of this argument is to extend NAEP to include a wider range of possible resources than are now included and to take some pains to insure that the resource indicators achieve a modicum of construct validity. For example, it would be extremely useful to field-test and validate student reports of multiple indicators of student social background including parental occupation, and to construct and validate a scale for cognitive stimulation in the classroom based on student reports. Linking NAEP data to indicators of neighborhood demographic characteristics such as poverty concentration, housing density, and ethnic composition would strengthen inference by allowing control for residential context. And it would be exciting to include with NAEP a survey of teachers in order to construct school-level indicators, based on teacher reports, of normative cohesion, expectations, collaboration, control, opportunities for learning, and school-level academic press. The availability of denser data at the level of the student, classroom, and school would provide a wider range of school resources than can now be studied, leading to a richer characterization of the association between student background and access to resources.

A promising avenue for future research is to develop more sophisticated models to explain variation in access to key resources. School district wealth, urban versus suburban versus rural location, school size, per pupil expenditures, and school social composition may shape the probability that resources will become available to a student; and studying such predictors may shed light on impediments to increasing access and identify new targets for intervention by policy. Our broad recommendation is that, as we assess student progress in subject-matter proficiency, we also assess the extent to which the education system provides resources that support such student progress.

References

Arnold, C. L., Kaufman, P. D., and Sedlacek, D. S. (1992). *School Effects on Educational Achievement in Mathematics and Science: 1985–1986 (Research and Development Report)*. U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., and Smith, T. A. (1996). *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Boston, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Berne, R. (1994). Educational Input and Outcome Inequities in New York State. In R. Berne and L. O. Picus (Eds.), *Outcome Equity in Education* (pp. 1–23). Thousand Oaks, CA: Corwin Press, Inc.

Bernstein, B. (1970). Education Cannot Compensate for Society. *New Society 387*: 344–347.

Bowles, S., and Gintis, H. (1976). *Schooling in Capitalist America*. New York: Basic Books.

Breslow, N., and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association 8:* 9–25.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods.* Newbury Park: Sage Publications.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. T. (1994). *An Introduction to HLM: Computer Program and Users' Guide. Version 2*. Chicago: Department of Education, University of Chicago.

Bryk, A. S., and Thum,Y. M. (1989). The Effects of High School Organization on Dropping Out: An Exploratory Investigation. *American Educational Research Journal 26*(3): 353–383.

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.

Finley, M. K. (1984). Teachers and Tracking in a Comprehensive High School. *Sociology of Education 5*: 233–243.

Fraser, E. (1959). *Home Environment and the School.* London: University of London Press.

Fuller, B. (1987). Raising School Quality in Developing Countries: What Investments Improve School Quality? *Review of Educational Research 57*: 255–291.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association 85*: 398–409.

Glass, G. V. (1976). Primary, Secondary, and Meta-analysis of Research. *Educational Researcher 5*: 3–8.

Johnson, E. G., Mazzeo, J., and Kline, D. L. (1993). *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kozol, J. (1991). *Savage Inequalities*. New York: Crown.

Lee, V. E., and Bryk, A. S. (1989). A Multilevel Model of the Social Distribution of High School Achievement. *Sociology of Education 62*: 172–192.

Lee, V. E., and Bryk, A. S. (1993). The Organization of Effective Secondary Schools. In L.Darling-Hammond (Ed.), *Review of Research in Education* (Ch. 5, pp. 171–267). Washington DC: American Educational Research Association.

Little, R. J., and Schenker, N. (1994). Missing Data. In G. Arminger, C. C. Clogg, and M. E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 39–75). New York: Plenum Press.

Medrich, E., and Griffith, J. (1992). *International Mathematics and Science Assessments: What Have We Learned.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Mislevy, R. J. (1992). Scaling Procedures in NAEP. Special Issue: National Assessment of Educational Progress. *Journal of Educational Statistics 17*(2): 131–154.

Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association 78*(381): 47-65.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1993, April). *NAEP 1992 Mathematics Report Card for the Nation and the States.* Princeton, NJ: Educational Testing Service.

Oakes, J. (1985). *Keeping Track: How Schools Structure Inequality*. New Haven, CT: Yale University Press.

Oakes, J. (1990). *Multiplying Inequalities: The Effects of Race, Social Class, and Ability Grouping on Access to Science and Mathematics Education*. Santa Monica, CA: RAND.

Page, R. N. (1990). The Lower Track Curriculum in a College-preparatory High School, *Curriculum Inquiry 20:* 249–282.

Peaker, G. F. (1967). *The Plowdon Children Four Years Later*. London: National Foundation for Education in England and Wales.

Raudenbush, S. W., and Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics 20*(4): 307–335.

Raudenbush, S. W. (1994). Random Effects Models. In H. Cooper and L. Hedges (Eds.), *The Handbook of Research Synthesis* (Chapter 20, pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W. (1988). Educational Applications of Hierarchical Linear Models: A Review. *Journal of Educational Statistics 13*: 85–116.

Raudenbush, S. W., Fotiu, R. P., and Cheong, Y. F. (in press). Synthesizing Results from the Trial State Assessment. *Journal of Educational and Behavioral Statistics.*

Raudenbush, S. W., Fotiu, R. P., and Cheong, Y. F. (1998). Inequality of Access to Educational Opportunity: A National Report Card for Eighth Grade Math. *Educational Evaluation and Policy Analysis 20*(4): 253–268.

Raudenbush, S. W., Rowan, B., and Cheong, Y. F. (1993). Higher Order Instructional Goals in Secondary Schools: Class, Teacher, and School Influences. *American Educational Research Journal 30*(3): 523–553.

Rist, R. (1970, August). Student Social Class and Teacher Expectations: The Self-ful filling Prophecy in Ghetto Education. *Harvard Educational Review 40*: 411–451.

Rosenbaum, J. E. (1976). *Making Inequality: The Hidden Curriculum of High School Tracking*. New York: Wiley.

Rosenthal, R., and Jacobson, L. (1968). *Pygmalion in the Classroom.* New York: Holt, Rinehart and Winston.

Rotberg, I. C. (1998, May 15). Interpretation of International Test Score Comparisons. *Science 280 (S366):* 1030–1031.

Rowan, B., Raudenbush, S. W., and Cheong, Y. F. (1993). Teaching as a Non-routine Task: Implications for the Organizational Design of Schools. *Educational Administration Quarterly 29*(4): 479–500.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., and Smith, A. (1979). *Fifteen Thousand Hours: Secondary Schools and Their Effects on Children*. Cambridge, MA: Harvard University Press.

Ryan, W. (1971). *Blaming the Victim*. New York: Pantheon.

U.S. Department of Education. National Center for Education Statistics. (1995). *The Condition of Education 1995*. Washington, DC: U.S. Government Printing Office.

Willms, J. (1992). *Monitoring School Performance: A Guide for Educators*. Washington, DC: Falmer Press.

Wolf, R. (1968). The Measurement of Environments. In A. Anastasi (Ed.), *Educational Testing Problems in Perspective*. Washington, DC: American Council on Education.

Zeger, S., Liang, K., and Albert, P. (1988). Models for Longitudinal Data: A Likelihood Approach. *Biometrics 44*: 1049–60.

Moving Educational Research Toward Scientific Consensus

David W. Grissmer
Ann Flanagan
RAND

Introduction

Educational research has been characterized, perhaps unfairly in recent years, by the inconsistency of its research results and by a lack of consensus across its broad and multidisciplinary research community (Wilson and Davis 1994; Saranson 1990). The broad purpose of this conference is to help determine how we can improve the consistency and accuracy of results in educational research so that we can build a base of knowledge widely accepted by this diverse research community and, more importantly, by teachers, principals, superintendents, and policymakers. To do so is a daunting task since education is one of the most complex topics addressed by social science. It is not surprising that progress in this direction has been slow, given both the broad interdisciplinary basis and the inherent complexity of learning.

We have proceeded with the hope that better nonexperimental data and more sophisticated model specifications and estimation techniques will eventually bring consensus. In this paper we will suggest that simply improving the kinds of nonexperimental data currently collected, along with the associated statistical methodologies, will never be sufficient to achieve the kind of scientific consensus needed to effectively guide educational policies.[1] Research shows that the effects we are trying to measure are quite complex. They often appear to be nonlinear, sensitive to contextual factors, moderately correlated among themselves, and subject to selection bias within families and schools. More-

[1] Support for this work came from the Center for Research on Educational Diversity and Excellence (CREDE), the NAEP Redesign Research Program, the NAEP Secondary Analysis Program, and Exxon Corporation.

over, some achievement effects are long-term, sustained long after an intervention has stopped; and some fade after a few years. These results may be only the tip of the iceberg, considering the complexity of the underlying developmental phenomena we are trying to understand.

This complexity places great demand on the quality of our data, the sophistication of our model specifications, and the accuracy of our estimation techniques. One interpretation of the wide variation in measurements of the effects of most factors affecting student achievement is that our data, model specifications, and estimation techniques do not yet reflect much of this inherent complexity. When results vary, it is difficult to determine why one set of results should be trusted over another, since practically every measurement makes different assumptions or uses different model specifications and estimation techniques. The wide variety of data quality, assumptions, and specifications may introduce enough bias and randomness to produce inconsistent effects across different data sets and model specifications. In this case the results should not be interpreted as “no effect,” but rather as inconclusive.

We suggest that three research approaches will be necessary to lead reliably to research consensus: increasing experimentation, building theories of educational process, and improving our nonexperimental analysis. Further, we believe that future data collection and research should be guided by a strategic plan built upon experimentation. Such a plan would provide the necessary data to build theories of educational process and improve our specifications of models used in nonexperimental analysis.

Experiments—if well designed, implemented, analyzed, and replicated—provide explanations that are as close to causal as possible in social science. Such experiments can provide the most accurate results for the effect of a particular variable in a given context. Experiments can also play another, and perhaps more important, role in social science research—namely, helping to validate model specifications for nonexperimental data. A key theme of this paper is that future experimentation and data collection need to be directed toward both the building of theories and the improvement of our assumptions in analyzing nonexperimental data. In the long run, policy analyses will largely be dependent on improving nonexperimental analysis since experiments can never be counted on to solve all the complex and contextual effects present in education.[2] Therefore, improving our confidence in the model specifications used with nonexperimental data is critical.

The major thrust of this paper is to suggest that building scientific consensus will require a coherent research strategy. This strategy must be built upon increasing experimentation, developing theories of educational process, and improving confidence in nonexperimental analysis, if we are to achieve research consensus. In this paper we focus initially on the broad lack of agreement relating to the effects of educational resources and social and educational policy on children. Thirty years of research with nonexperimental data have led to almost no consensus on these important policy issues. We then focus on a narrower question, namely, the impact of resources on educational outcomes, particularly student achievement. This situation presents an interesting case study where a consensus based on the results of nonexperimental data once existed, only to be challenged recently by new experimental and nonexperimental research.

We use the Tennessee class size experiment results to illustrate the process of deriving "rules" for model specification used in nonexperimental data involving class size. We then illustrate the process of building theories of educational process related to class size effects and describe the role of such theories in building stronger consensus. Finally, we specifically focus on implications for the National Assessment of Educational Progress (NAEP) and other data collections and more generally suggest directions for future research and development (R&D) efforts to build a more solid foundation of knowledge for educational policymaking.

Children's Well-Being: The Ongoing Debate

Federal, state, and local governments spend approximately $500 billion per year in social, educational, and criminal justice expenditures on the nation's

[2] Large-scale experiments such as the Tennessee class size experiment can be costly and take considerable time to plan, implement, and analyze. While more experimentation seems essential to making progress in educational research, educational research will probably never follow health research, where trials are needed for every new intervention before implementation.

children and youth (Office of Science and Technology Policy 1997).[3] The amount spent on children appears to have increased substantially over time (Fuchs and Rekliss 1992), although there is debate about the magnitude of the real increase in spending. Thus, an important set of public policy questions is associated with how effective this increased spending has been at improving the well-being of our children. Besides increased investment, there have been significant changes in families, communities, and schools that would be expected to affect children's outcomes.

There is little scholarly consensus about the effects of expenditures on children or the effects from changing families, communities, and schools. For instance, scholars disagree about the impact of the War on Poverty and expanded social welfare programs (Herrnstein and Murray 1994; Jencks 1992); they also disagree on whether increased school resources have raised student achievement levels (Burtless 1996; Ladd 1996a). There is disagreement about the way communities have changed for black families (Wilson 1987; Jencks 1992) and whether the net effect on children of recent changes in the family has been positive or negative (Cherlin 1988; Zill and Rogers 1988; Fuchs and Rekliss 1992; Popenoe 1993; Stacey 1993; Haveman and Wolfe 1994, 1995; Grissmer et al. 1994). There is more agreement about the effects of desegregation, although some dispute remains (Wells and Crain 1994; Schofield 1995; Armor 1995; Orfield and Eaton 1996). Finally, many small-scale, intensive early childhood programs appear to produce significant short- and long-term effects, but there is disagreement about large-scale programs—how large the effects from attending kindergarten and preschool are and how long these effects last (Barnett 1995; Karweit 1989). Recent evidence suggests that the cost-effectiveness of early childhood programs can depend critically on the characteristics of the targeted group, with significant net fiscal returns for some groups, but not others (Karoly et al. 1998).

[3] This estimate does not include the foregone taxes for deductions for children and day care. Besides public sector spending on children, approximately $560 billion is spent in the private sector on children, bringing the average public and private spending per child to approximately $15,000 annually. This amount is estimated assuming the cost of raising a child to age 18 to be approximately $150,000, with approximately 70 million individuals between the ages 0–18. Thus, annual expenditures are $150,000 x 70,000,000/18 = $560 billion. See United States Department of Agriculture (1997) for estimates of the cost of raising children.

Despite the lack of consensus among the educational research community, dramatic changes are being proposed and are occurring in both social and educational policies, based on perceptions that past policies have failed. For instance, much of the movement toward more fundamental reform of public schools arises from perceptions that massive increases in resources in grades K–12 education over the last 25 years have resulted in declining—or at best stable—student achievement (as measured by scores on the Scholastic Achievement Test [SAT] and NAEP scores) and that schools have particularly failed minority students. If so, a solid case could be made for restructuring school governance and incentive structures so that more effective utilization of resources might possibly occur (Hanushek 1994; Hanushek and Jorgenson 1996). However, new research is challenging this once widely accepted conclusion.

A Shifting Consensus: The Effects of Educational Resources[4]

Until the early to mid-1990s, the dominant research position among social scientists was that school resources had little impact on student achievement. This counterintuitive view dated from the "Coleman report" (Coleman et al. 1966). Influential reviews by Eric Hanushek (1989, 1994, 1996, 1999) also argued that evidence from over 300 empirical measurements provided no consistent evidence that increases in school resources raised achievement scores. It was suggested that a key reason for inefficiency in public schools was a lack of incentives (Hanushek and Jorgenson 1996).

However, it would not be surprising that some money was spent inefficiently, given that no definitive results emerged from educational research that could guide policymakers. At worst—if past resources can be shown to have had no effect on achievement—this finding can simply indicate the lack of guidance by good R&D. The lack of a critical level of R&D funding and critical mass of high quality research may provide an explanation for inefficiency just as persuasive as the lack of incentives (Wilson and Davis 1994).

[4] The early sections of this paper draw heavily from four recent papers—Grissmer, Flanagan, and Williamson (1998a); Grissmer et al. (1998); and Grissmer, Flanagan, and Williamson (1998b); and Grissmer et al. (forthcoming). We have quoted liberally from these papers without quotation marks.

Hanushek's original reviews did not group studies using the quality of data and specifications, type of intervention, or student or grade level (1989, 1994). However, Hanushek refined his reviews, focusing on effects from per pupil expenditure and pupil/teacher ratio reductions and disaggregating studies by grade level, level of aggregation, and model specifications (Hanushek 1996, 1999). These later reviews still indicated that subsets of studies provide positive and negative coefficients in about equal numbers. One focus was on studies using a production function framework where the previous year's test scores were used as controls. These models were judged by many to be the most likely to avoid bias. These models also showed balanced numbers of positive and negative coefficients. These results strengthened the conclusion that the nonexperimental evidence supported little effect from class size reductions or additional expenditures.

Subsequent literature reviews questioned the selection criteria used in Hanushek's reviews to choose studies for inclusion and the assignment of equal weight to all measurements from the included studies. Two subsequent literature reviews (Hedges, Laine, and Greenwald 1994; Krueger 1999a) used the same studies included in Hanushek's reviews, but came to different conclusions. One study used meta-analytic statistical techniques for combining the measurements, which do not weigh each measurement equally (Hedges, Laine, and Greenwald 1994). Explicit statistical tests were made for several variables for the hypotheses that the results support a mean positive coefficient and reject a mean negative coefficient. The results concluded that, for most resource variables, the results supported a positive relationship between resources and outcomes. In particular, per pupil expenditures and teacher experience provided the most consistent positive effects, with pupil/teacher ratio, teacher salary and teacher education having much weaker effects.

A more recent literature review using the same studies included in Hanushek's reviews also concludes that a positive relationship exists between resources and outcomes (Krueger 1999a). This review criticizes the inclusion and equal weighting of multiple measurements from single published studies. Some studies provided as many as 24 separate measurements due to the presentation of sets of results for many subgroups. Since the average sample size will decline as subgroups increase, many of the measurements lacked the statistical power to detect policy-significant effects; and thus many insignificant coefficients might be expected. Since the presentation of results for subgroups is not done uniformly across studies, and may even be dependent on the results obtained, Krueger (1999a) reanalyzes the data to determine if the inclusion of multiple measurements significantly affects the conclusions reached. His analysis concludes that the inclusion of multiple measurements is a significant factor in explaining the original conclusions, and that less weight placed on these multiple measurements would lead to support for a positive relationship between higher per pupil expenditures and lower pupil/teacher ratio and outcomes.

A more comprehensive review of the literature prior to 1990 used meta-analytic statistical comparison techniques, but searched a wider literature and imposed different quality controls (Greenwald, Hedges, and Laine 1996). All the included studies used achievement as the dependent variable and measurements at the individual or school level only. The resulting set of measurements utilized in the study included many measurements that were not included in Hanushek's studies and rejection of about two-thirds of the measurements included in Hanushek's reviews.

The conclusions analyzing the set of coefficients from six variables (per pupil expenditure, teacher ability, teacher education, teacher experience, pupil/teacher ratio, school size) supported statistically the hypothesis that the median coefficients from previous studies showed positive relationships between resource variables and achievement. However, the variance in coefficients for each variable across studies was very large. Extreme outliers appeared to be a problem for some variables, and the coefficients across studies appeared to have little central tendency indicating the presence of nonrandom errors.

This review also reported results for measurements using different model specifications (longitudinal, quasi-longitudinal and cross-sectional).[5] The results showed that median coefficients changed dramatically for most variables across specifications, with no recognizable pattern. Although few studies had what were considered to have superior specifications (longitudinal studies), the median coefficients for these models were negative for per pupil expenditure, teacher education, pupil/teacher ratio, and school size. When the median coefficients of studies having quasi-longitudinal studies were compared to coefficients from the entire sample, results were similar for four variables, but differed for the remaining two variables by factors ranging from 2 to 20. In the

[5] Longitudinal studies were defined as those having a pretest control score, and quasi-longitudinal was defined as having some earlier performance-based measure as a control. Cross-sectional studies merely had SES-type variables included as controls.

case of teacher salary, these studies provided a median coefficient indicating that a $1,000 salary increase could boost achievement by over one-half standard deviation.

This review utilized better screening criteria and better statistical tests to conclude that the overall evidence supported positive effects from additional resources. However, the large variance in coefficients and the sensitivity of the median coefficients to which studies were included provided little confidence that the literature could be used to estimate reliable coefficients. In particular, models thought to have superior specifications provided no more consistent results and sometimes provided noncredible estimates.

Besides the argument from literature reviews, Hanushek made another argument that seemed consistent with his conclusions. Measured in constant dollars, expenditures per pupil doubled between the late 1960s and the early 1990s; however, NAEP scores at age 9, 13, and 17 showed no dramatic improvement in average reading or math skills during this period. We address this argument next.

Interpreting NAEP Score Trends

Achievement scores are a particularly good measure of the changing environment for our children since research has shown that achievement reflects the combined influence of families, communities, and schools. Significant changes in the quality of our families, schools, and communities should be reflected on achievement trends that are best measured by NAEP (Cambell et al. 1996; Miller, Nelson, and Naifeh 1995; Mullis et al. 1993; Reese et al. 1997).

The NAEP achievement scores collected from 9-, 13-, and 17-year-olds since 1969 are the only nationally representative achievement scores available. The primary purpose of NAEP has been to simply monitor the achievement of American students; however, NAEP scores are increasingly being used to evaluate the effects on youth from the dramatic changes in families, communities, and schools, and from our nation’s educational and social policies—changes that have taken place since the late 1960s. These changes include the following:

- National efforts to equalize opportunity and reduce poverty that began in the mid-1960s and continued or expanded in subsequent decades. These efforts included federally funded preschools (e.g., Head Start), compensatory funding of elementary schools with large numbers of low-income students, desegregation of schools, affirmative action in college and professional school admissions, and expanded social welfare programs for poor families.
- Changes in school attendance and school changes that were not primarily designed to equalize opportunity. These changes included increased early schooling, greater per pupil expenditures, smaller classes, significant changes in the characteristics of teachers, and systemic reform initiatives.
- Changes in families and communities that may have been somewhat influenced by efforts to equalize opportunity and reduce poverty but that occurred mainly for other reasons. Specifically, parents acquired more formal education, more children lived with only one parent, more children had only one or two siblings, and the proportion of children living in poverty rose. At the same time, poor blacks concentrated more in inner cities, while the more affluent blacks moved to the suburbs.

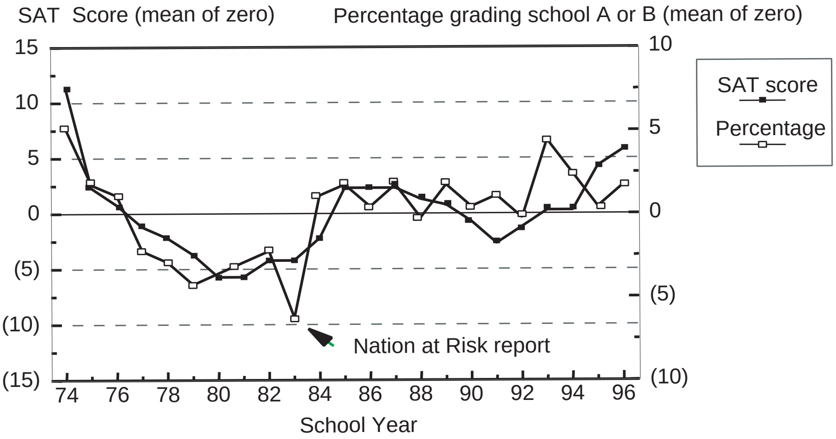
The 17-year-olds tested by NAEP in 1971 would have grown up in families and communities and attended schools largely unaffected by the changes cited above. However, those recently tested would have lived their entire lives in families, communities, and schools reshaped by these policies. It would be hard to take a position about the quality of our families, communities, and schools and the effectiveness of social and educational policies that would be inconsistent with the trends in the NAEP data.

Until recently, the NAEP scores were used only peripherally to address these kinds of questions, partly because the more widely recognized (but fatally flawed) SAT scores were used whenever test scores entered the public debate. One reason that SAT scores are used effectively in public debate is that the public appears to base its assessment of the quality of American schools on SAT scores (Grissmer forthcoming). Figure 1 shows the results of an annual public opinion poll that asks adults to grade the nation’s schools. The percentage of adults giving schools an “A” or a “B” is graphed against changes in annual average SAT scores.[6] The data show that public opinion appears to follow the SAT trends.

The well-known flaws in the SAT scores for monitoring national achievement trends result from their self-selected sample (Advisory Group on the Scholastic Aptitude Test Score Decline 1977; Koretz 1986, 1987; Rock 1987; Grissmer et al. 1994). The scores are biased downward, not only because of an increasing percentage of students taking the test but also because the students making the largest achievement gains from 1970 to 1990—minority and disadvantaged students—are largely missed by the SAT because they do not go to college. Ironically, if K–12 education improves, allowing more children to attend college, the SAT scores will decline. Thus, SAT scores are probably a perverse indicator of K–12 school quality.

The research community switched to analyzing NAEP data in isolated studies dating from the mid-1980s. A steady stream of analyses from the late 1980s drawn from the NAEP data developed into more detailed analyses using

Figure 1. Comparing the trends in SAT scores with percentage of adults giving schools a grade of "A" or "B"



[6] The graph normalizes both variables to a mean of 0. The regression fit for the equation, School grade = a + b (Average SAT score), gives b = .79 (t = 5.2), R-Squared = .56.

new methodologies from the mid-1990s.[7] Early work took note of the large gains in black scores and the very small gains in white scores, along with the resulting convergence of the black-white test score gap. The contrast with falling SAT scores was noted. However, familiarity with this earlier work—buttressed by the National Research Council (1989)—seemed to remain confined to a small group of researchers, and declining SAT scores remained the dominant influence among both the public and the research community.[8]

Starting in the early 1990s, analyses of the NAEP data began to provide more detail about differences in trends among black, Hispanic, and white students; differences in trends for lower- and higher-scoring students; differences by age; and particularly differences by entry cohorts. The analyses also attempted to explain the trends and the convergence in the black-white test score gap.

Across ages and subjects, the largest gains in scores occurred for black students; but significant gains were registered by Hispanic students and lower-scoring white students, with small gains or none registered by average and higher-scoring white students (Hedges and Nowell 1998; Hauser 1998; Grissmer et al. 1994, 1998; Grissmer, Flanagan, and Williamson 1998a). These studies also noted the evidence that black gains were largely confined to a group of about 10 cohorts born in the mid-1960s to the mid-1970s and entering school around 1970 to 1980. For later cohorts, black scores and the black-white achievement gap have—for most age groups and subjects—remained stable or declined.

The most striking feature of the NAEP results for blacks is the size of adolescents' gains for cohorts entering from 1968–1972 to 1976–1980. These

[7] See Hauser (1998) for a history of utilizing NAEP scores from 1984 to 1992. This period included work by Jones (1984); Koretz (1986, 1987); National Research Council (1989); Linn and Dunbar (1990); and Smith and O'Day (1991). See Rothstein (1998) for a long-term history of achievement that extends through 1997. This paper draws from all of these studies.

[8] This phenomenon points to a second problem in attaining consensus in the educational research community. While small groups of researchers with in-depth knowledge in a subject may find consensus, it is quite another problem for this information to be disseminated, accepted broadly, and commonly cited in most research. The diverse set of journals and disciplinary boundaries make it difficult for narrow consensus to become broad consensus.

gains were 0.6 standard deviation averaged across reading and math. Such large gains for very large national populations over such short time periods are rare, if not unprecedented. Scores on IQ tests given to national populations seem to have increased gradually and persistently throughout the 20th century, both in the United States and elsewhere (Flynn 1987; Neisser 1998). But no evidence exists in these data involving large populations showing gains even close to the magnitude of the gains made by black student cohorts over a 10-year period.

Even in intensive programs explicitly aimed at raising test scores, it is unusual to obtain gains of this magnitude. Early childhood interventions are widely thought to have the largest potential effect on academic achievement, partly because of their influence on brain development. Yet only a handful of "model" programs have reported gains as large as half a standard deviation (Barnett 1995). These programs were very small-scale programs with intensive levels of intervention. Even when early childhood programs produce large initial gains, the effects usually fade at later ages. Among blacks who entered school between roughly 1968 and 1978, in contrast, the gains were very large among older students and were not confined to small samples, but occurred nationwide.

Beginning in the mid-1990s, finding the likely causes of these gains became the focus of research. Part of the quest was to determine whether the dramatic changes that occurred in families during this period could explain the gains. Utilizing data from several sources (Current Population Survey [CPS], the National Longitudinal Survey of Youth [NLSY], and the National Education Longitudinal Study [NELS]), one study developed a new methodology to estimate the size of the net expected gains from changes in eight key family characteristics for 13- to 17-year-old test-takers from 1970–90 (Grissmer et al. 1994). The analysis required several assumptions—one concerning the stability of family coefficients in achievement equations over time.[9] The results of the analysis indicated that changes in the family would predict small positive gains in scores for all racial-ethnic groups and that these gains could account for the smaller score gains among whites but could explain only about one-quarter of the minority gains.

[9] Evidence from Hedges and Nowell (1998) and Cook and Evans (1997) appears to support fairly stable family coefficients over time.

Another analysis using NAEP individual level data also concluded that family effects could account for only one-quarter or so of black gains (Cook and Evans 1997). This analysis relied on student-reported family characteristics collected with the NAEP, but utilized a methodology newly imported from labor economics to attempt to partition the gains into those related to family changes, changes in family structural characteristics, and those due to changes between and within schools. If effects from changing family characteristics are small, the likely remaining hypothesis for the black score gains is school-related, community-related, or related to yet unmeasured family characteristics.

Jencks and Phillips (1998) summarized research efforts focusing on the black-white test score gap. Their book brought together a diverse set of scholars to try to determine where consensus can be achieved on this topic and where and what kind of additional research is needed.[10] Three analyses reported in the book look at the convergence and possible divergence of the black-white score gap for cohorts born as early as 1950 (Hedges and Nowell 1998; Phillips, Crouse, and Ralph 1998; Grissmer, Flanagan, and Williamson 1998a). Two of the studies utilize NAEP data as well as achievement and survey data from other studies. All agree that significant narrowing occurred for cohorts born prior to about 1978—but no further narrowing occurred for later cohorts.

Although the black-white gap for reading actually widened, Phillips, Crouse, and Ralph (1998) concluded that the widening is not statistically significant. Hedges and Nowell (1998) and Grissmer, Flanagan, and Williamson (1998a) provided evidence that family changes may explain a part of the narrowing. Further, Grissmer, Flanagan, and Williamson (1998a) observed that the timing of the black gains by age group and region suggested two major hypotheses for the gains. The first hypothesis was based on changes in schooling—changing pupil/teacher ratios and class sizes, changing teacher characteristics, and changing curricula. Changing pupil/teacher ratios emerged

[10] In the process of achieving consensus, support for a continuing series of books dedicated entirely to exploring the most important questions in education seems crucial. Besides Jencks and Phillips (1998), Ladd (1996a) and Burtless (1996) are also good examples. In these latter books, the consensus might be characterized more by what is not known than what is known.

as a viable, but not completely satisfactory, explanation in other analyses (Krueger 1998; Ferguson 1998).[11]

A second explanation emerged, more closely related to the changes engendered by the Civil Rights movement and the War on Poverty. Such changes could have direct effects related to school desegregation—particularly in the South—and indirect effects caused by the perceived shift in the motivation for and attitudes toward education of black parents and students stemming from better opportunities for future schooling and jobs. An additional possible shift from these efforts could have occurred in the behavior and attitudes of teachers of black students that resulted in increased attention and resources. The timing of the black gains by age coincides with the broad-scale implementation of such efforts, if the assumption is made that most of the effects would occur only if students experienced these changes from the early grades forward. The large gains for minority and disadvantaged students, as well as the smaller gains (or lack of gain) among average and higher-scoring white students, pose a challenge to the thesis that the increased spending in education and social programs aimed at these students was ineffective.

Analysis of NAEP scores appears to be central to the debates about changes in American families and schools, policies providing equal opportunity in education, and the best way to spend investments in education and children. The effective absence of these scores from these national debates has allowed many widespread beliefs to proliferate that seem to be at odds with the NAEP results. The NAEP data do not suggest that families have deteriorated since 1970. Nor do they suggest that schools have spent money inefficiently or that social and educational policies aimed at helping minorities have failed.

Instead, they suggest that family environments changed in positive ways from 1970 to 1996, that the implementation of the policies associated with the Civil Rights movement and the War on Poverty may be a viable explanation for large gains in black scores, and that certain changes in our schools and curriculum are consistent with NAEP score gains. While the NAEP scores

[11] The timing of pupil/teacher ratio changes would suggest that score gains should have started earlier and would affect white scores as well—leading to overpredicted white gains. Further research to determine whether class size for black students fell more than for white students might help reduce the overprediction of white score gains. This overprediction would also be addressed if class size reductions were small or nonexistent for more advantaged white students.

alone cannot reject the beliefs about deteriorating families and schools and the ineffectiveness of social and educational policies, the advocates of such beliefs must provide an explanation for NAEP scores consistent with their positions. The NAEP scores from 1971 to 1988 generally support a more positive picture of our families, schools, and public policies; however, trends in black achievement since 1988 to 1990 have been more discouraging, and it is critical to understand why these reversals have occurred.

Trends in School Resources

Research on NAEP scores shows that the increases were negligible only for the higher-scoring white population, but substantial for black, Hispanic, and lower-scoring white students. A second line of research using new data and new methods of estimating "real" per pupil expenditures over time shows that resource growth tended to occur where achievement gains were made (Rothstein and Miles 1995).

A new method of deflating school expenditures, taking account of the labor intensity of schools, showed that resources did not come close to doubling as had been indicated by the commonly used Consumer Price Index (CPI). Use of more appropriate indices for adjustment of educational expenditures reflecting their labor intensity provides much lower estimates of real growth (Rothstein and Miles 1995; Ladd 1996b).

Moreover, the new method—developed to assign school expenditures to programmatic categories that could distinguish spending on different types of students—showed that even this smaller increase overestimates the additional resources available to boost achievement scores for regular students. A large part of the smaller estimated increase went for students with learning disabilities, many of whom are not tested.[12] Another part also went for other socially

[12] There is agreement that a disproportionate fraction of the expenditure increase during the NAEP period was directed toward special education (Lankford and Wyckoff 1996; Hanushek and Rivkin 1997). Hanushek and Rivkin estimated that about a third of the increase between 1980 and 1990 was related to special education. NAEP typically excludes about 5 percent of students who have serious learning disabilities. However, special education counts increased from about 8 percent of all students in 1976–77 to about 12 percent in 1993–94. These figures imply that 7 percent of students taking the NAEP tests were receiving special education resources in 1994, compared to 3 percent in 1976–77. This percentage is too small to have much effect on NAEP trends, but it should in principle have had a small positive effect.

desirable objectives that are only indirectly related to academic achievement. Taking into account better cost indices, and including only the spending that would have been directed at increasing achievement scores, Rothstein and Miles (1995) concluded that the real increase in per pupil spending on regular students was closer to 30 than to 100 percent.

These smaller additional expenditures for regular students are mainly accounted for by lower pupil/teacher ratios, increased teacher salaries due to more experienced and educated teachers, and compensatory programs that would be expected to benefit minority and lower income students (Rothstein and Miles 1995; Hanushek and Rivkin 1997). The key issue then becomes whether these resource increases can plausibly explain any part of the pattern of large black gains and the absence of white gains unaccounted for by family changes. This pattern might be explained if black students received disproportionate shares of the additional resources or if black students benefited more than white students due to similar increases in resources.[13]

The Tennessee Experiment

Important new evidence for challenging the view that money doesn't matter comes from a large-scale experiment in Tennessee on the effects of class size. The Tennessee experiment in education was largely ignored for several years by the wider research community, and only recently has been reanalyzed and given its deserved prominence (Ritter and Boruch 1999). This experimental research suggests that reductions in class size may, in fact, have more impact on disadvantaged and minority students than on white students. A quasi-experiment in Wisconsin that varied student/teacher ratio also provided new evidence (Molnar et al. 1999).

The first experimental evidence on the effect of major educational variables came from a Tennessee study on the effects of class size (Word, Johnston, and Bain 1990; Finn and Achilles 1990; Mosteller 1995). About 79 schools in

[13] A number of policies sought to shift resources toward minority or low-income students during these years, including federal compensatory funding based on the percentage of children in poverty, school desegregation, and court-directed or legislative changes in state funding formulas toward minority and low-income school districts. However, other factors operated over this time period that could have increased funding for middle- and upper-income children as well. It is still unclear whether the net effect has been to disproportionately shift resources toward minority and lower-income children.

Tennessee randomly assigned about 6,000 kindergarten students to class sizes of approximately 15 or 23 students, and largely maintained their class size through third grade. Additional students entering each school at first, second, and third grade were also randomly assigned to these classes making the entire experimental sample approximately 12,000. After third grade, all students were returned to standard, large-size classes through eighth grade. The students in the experiment were disproportionately minority and disadvantaged—33 percent were minority, and over 50 percent were eligible for free lunch.

Analysis of the experimental data shows statistically significant, positive effects from smaller classes at the end of each grade from K–8 in every subject tested (Finn and Achilles 1999; Krueger 1999b; Nye, Hedges, and Konstantopoulos 1999; Nye, Hedges, and Konstantopoulos forthcoming). The magnitude of results varies depending on student characteristics and the number of grades in small classes. Measurement of effect sizes from four years in small classes at third grade varies from 0.25 to 0.4 standard deviation (Krueger 1999b; Nye, Hedges, and Konstantopoulos forthcoming). The current measurement of long-term effects at eighth grade show sustained effects of approximately 0.4 standard deviation for those in small classes all four years, but little sustained effect for those in smaller classes one or two years (Nye, Hedges, and Konstantopoulos 1999). Short-term effects are significantly larger for black students and somewhat larger for those receiving free lunches.[14]

Questions were raised whether the inevitable departures from experimental design that occur in implementing the experiment biased the results (Krueger 1999b; Hanushek 1999). These problems included attrition from the samples, leakage of students between small and large classes, possible nonrandomness of teacher assignments, and schooling effects. Recent analysis has addressed these problems without finding any significant bias in the results (Krueger 1999b; Nye, Hedges, and Konstantopoulos 1999; Nye, Hedges, and Konstantopoulos forthcoming; Grissmer 1999). It is possible for further analysis to find a flaw in the experiment that significantly affects the results, but extensive analysis to date has eliminated most of the potential problems.

[14] Long-term effects have not been reported by student characteristics. Following the experiment, Tennessee also cut class sizes to about 14 students per class in 17 school districts with the lowest family income. Comparisons with other districts and within districts before and after the change showed even larger gains of 0.35 to 0.5 standard deviations (Word, Johnston, and Bain 1994); Mosteller 1995). Thus the evidence here suggests that class size effects may grow for the most disadvantaged students.

The Wisconsin SAGE (Student Achievement Guarantee in Education) quasi-experimental study differed in several important ways from the Tennessee STAR experiment (Molnar et al. 1999). In the SAGE study, only schools with very high proportions of free-lunch students were eligible for inclusion. Assignments were not randomized within schools, but rather a preselected control group of students from different schools was matched as a group to the students in treatment schools. The treatment is more accurately characterized as pupil/teacher ratio reduction since a significant number of schools chose two teachers in a large class rather than one teacher in a small class. The size of the reduction in pupil/teacher ratio was slightly larger than the class size reductions in Tennessee.

There were about 1,600 students in the small pupil/teacher treatment group in Wisconsin, compared to approximately 2,000 students in small classes in Tennessee. However, the size of control groups differed markedly—around 1,300 students in Wisconsin and around 4,000 in Tennessee, if both regular and regular-with-aide classes are combined. The SAGE sample had approximately 50 percent minority students with almost 70 percent eligible for free or reduced price lunch.

The results from the Wisconsin study for two consecutive first grade classes show statistically significant effects on achievement in all subjects (Molnar et al. 1999). The effect sizes in the first grade are in the range of 0.1–0.3 standard deviations. The lower estimates between 0.1–0.2 occur in regression estimates, while the raw effects and hierarchical linear modeling (HLM) estimates are in the 0.2–0.3 range. While the estimates seem consistent with the Tennessee study at first grade, more analysis is needed before the results can be compared.

Learning From the Tennessee Experiment about Model Specification

One of the problems with nonexperimental data analysis is that the research community usually fails to completely list the assumptions that are required in any analysis to make the analysis equivalent to experimental data.[15] Such listing of assumptions would make much more explicit the wide gap that exists between experimental and nonexperimental data analysis.

Partly because there have been so few experiments in education, we have not paid much attention to their potentially critical role in shaping theories about education, helping to correctly specify variables and models using nonexperimental data, and specifying what data we should collect. If applied to reliable experimental data, models used to estimate nonexperimental data should be able to duplicate the experimental results. Krueger (1999b) suggests that production functions with previous year's score do not duplicate the Tennessee effects except in the first year of smaller classes. This larger first-year effect has been interpreted as a socialization effect.

The Tennessee results suggest several further specification issues. First, schooling variables in one grade can influence achievement at *all* later grades, so conditions in all previous years of schooling need to be present in specifications. Second, a pretest score cannot control for previous schooling characteristics. *The Tennessee results suggest that two students can have similar pretest scores, similar schooling conditions during a grade, and emerge with different posttest grades influenced by different earlier schooling conditions.* For instance, despite having similar schooling conditions in grades 4–8, relative changes in achievement occurred in those grades for those having one to two or three to four years in small classes in K–3. Another way of stating this analytically is that effect sizes at a given grade can depend on interactions between this year's schooling characteristics and all previous years' characteristics.

The production function framework using pretest controls assumes that any differences in pre- and posttests are captured by changed inputs during the period. The Tennessee results suggest that coefficients of such specifications are *un-interpretable from a policy perspective* since the effect of a change in resources during a period cannot fully be known until past and future school-

[15] An excellent counterexample is Ferguson and Ladd (1996), which starts to describe the conditions for a "gold standard" model and provides one of the most complete listings of assumptions of any economic analysis. Raudenbush and Wilms (1995) and Raudenbush (1994) also carefully outline the statistical assumptions in two kinds of models used in education. See also Heckman, Layne-Farrar, and Todd (1996) for an analysis that tests and provides evidence of the weakness of the assumptions inherent in a certain kind of model linking educational outcomes to educational resources.

ing conditions are specified. Thus the answer to the question of whether a smaller class size in second grade had an effect cannot be known until later grades, and the answer will depend on what the class sizes were in previous and higher grades.

Another interpretation of the Tennessee data is possible—namely, that reduced class size is a multiyear effect whose precise pattern is dependent on duration. Being in a small class not only raises short-term achievement in the current year, but also has an effect in succeeding years. Then the effect in first grade consists of a residual effect from kindergarten plus an independent first grade effect. The second grade effect is the sum of the residuals from kindergarten and first grade, plus an independent second grade effect. This explanation would account for the increasing effect with more years in small classes in the K–3 years—but would also account for the pattern after return to larger classes after third grade. Clearly there is residual, and continuing, effect from having attended smaller classes in grades K–3. However, the permanence of the effect depends on duration, indicating the effects are not simply additive.

Conceptually this makes the effect of class size reductions resemble a human "capital" input that can change output over all future periods, and models specifying the effects of capital investments may be more appropriate.[16] Production functions generally assume constant levels of capital, but children's human "capital" is probably constantly changing and growing.

From the standpoint of child development, these results are consistent with the concepts of risk and resiliency in children (Masten 1994; Rutter 1988). Children 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 year or two years do not.

Few, if any, previous studies have included variables for prior years' school characteristics from elementary school. At the individual level, virtually no longitudinal data from kindergarten were available. At more aggregate district

[16] Production functions are typically applied to model complete growth cycles in agriculture or other areas. We have tried to apply it to much smaller increments of growth in children by using pre- and post-test results. Production functions may have done less well in earlier studies predicting weekly plant growth as oppused to the complete cycle of growth over a season.

and state levels, data are usually available describing average characteristics for earlier years, but were probably seldom used.

Since most data sets at the individual level, such as NELS, do not contain the previous year's history for grades K–8, they cannot be used to estimate class size effects under this hypothesis.[17] Probably most previous measurements at the individual level have not had such data, and this might explain the downward bias in results. However, models using aggregate data have more of a chance at being able to include previous history—on average—for students in the sample. For instance, at a school district level, data would be available on class sizes in previous years. If no in-migration and out-migration occurs, then the average class size for district students can be determined for previous years. Migration will weaken the validity of these estimates, which means that higher levels of aggregation (state level data) will likely capture more accurately the historical class size for students in the aggregate sample.

The usual tendency for researchers is to trust the results of individual level analysis more than those of aggregate level analysis. This trust arises from several factors: larger sample size, more variance in variables, and sometimes more detailed family data. However, individual level analysis is to be preferred over aggregate level only if the quality of variables is equivalent. If aggregate level data can better capture accurate historical information, then these estimates may produce better results. Another implication is that our data collection efforts should focus on longitudinal data from early years. Use of longitudinal data beginning at or prior to school entry can sort out some of the specification problems that may exist in previous analyses.

There are two new sources of such longitudinal data that will include school, teacher, and family characteristics and achievement data. First, there are the newly emerging longitudinal state databases that link student achievement across years. Such data have very large sample sizes, and linkages are possible with teacher data and school characteristics. These data will be better able to address some of the potential specification issues involving dependence

[17] The current year's class size will work if it is highly correlated with all past years' class sizes. However, at the individual level it seems likely that the random elements that determine year-to-year class size—including in-migration and out-migration and decisions when to create additional classes—would not make this year's class size a particularly good predictor of previous years' sizes, particularly over many grades. However, this correlation should be explored.

of later achievement on the previous year's class size as well as thresholds and interactions with teacher characteristics. It may also be possible to determine class size effects in later grades as well as in early grades. The second source will be the Early Childhood Longitudinal Study (ECLS) funded by the U.S. Department of Education, which will collect very detailed data on children, their families, and their schools. These data will be much richer in variables, but much smaller in sample size than the state data sets.

A Weak Test of the Hypothesis Using State NAEP Data

Analysis of state NAEP scores is providing preliminary supportive evidence that certain state policies *do* matter in improving scores, that minority and disadvantaged students show the most gain from increased resources, and that the distribution of key resources is inequitable (Grissmer et al. forthcoming; see also Raudenbush in this volume).

We have used the state NAEP data for the seven reading and math tests given between 1990 and 1996 at the fourth or eighth grade level to test two hypotheses:

- whether aggregate state results provide estimates of pupil/teacher ratio that are in reasonable agreement with the Tennessee class size effects; and
- whether these results change when we utilize a pupil/teacher ratio variable incorporating only the current year of the NAEP test vs. the average of all previous years in school.

Estimates have been made using the 271 average state scores in equations controlling for the effects of different family and demographic characteristics of students across states (Grissmer et al. forthcoming). We have utilized three different ways of controlling for family characteristics at the state level. We have supplemented the NAEP family characteristics with Census data to derive more accurate family variables than those provided by NAEP (Grissmer et al. forthcoming). We have also utilized SES-like variables derived from the NELS and Census data. We found little difference in results across these family measures.

We used a random-effects model and estimated with the generalized linear estimator with exchangeable correlation structure, which takes account of the lack of independence of state observations across tests (produces robust standard errors), the unbalanced panels, and heteroskedascity. We also have made estimates with generalized least squares and maximum likelihood, achieving almost identical results.

In the equations linking average state scores to family and state educational characteristics, we included four educational variables that account for 95 percent of the variance in per pupil spending across states. These variables are average teacher salary, pupil/teacher ratio, teacher-reported adequacy of resources, and percentage of students in a state in public prekindergarten.[18] We found the expected signs and statistical significance for pupil/teacher ratio, teacher-reported resources, and prekindergarten participation. We found insignificant results for teacher salary.

The pupil/teacher ratio effect in this model would predict a rise of about 0.14 standard deviation for reduction of eight pupils per class (approximately the size of the Tennessee class size reductions). This effect is markedly smaller than the reported Tennessee class size effect of around 0.20–0.25. However, if we include in our models an interaction term allowing larger pupil/teacher effects for states with more disadvantaged students, we find markedly larger effects for states having more disadvantaged students. The Tennessee experimental sample contained a disproportionate percentage of minority and free lunch students, compared to all Tennessee students (Krueger 1999b). If we take into account the characteristics of the Tennessee sample and the interaction effect, the equations would predict a class size effect for the Tennessee sample that agrees with the actual effect.

We have tested whether results for pupil/teacher ratio differed in our data set when the variables were defined using pupil/teacher averages during time in school vs. pupil/teacher value in the year of the test only. We use the state average pupil/teacher ratio during all years in school, the average during grades 1 through 4, and the value in the year of the test. The estimates for these vari-

[18] We used a pupil/teacher variable rather than class size, since data were only available by year by state for the pupil/teacher ratio. While the two are highly correlated, one cannot necessarily assume that reductions in pupil/teacher ratio and class size would produce the same effects.

Table 1. Comparing Three Pupil-Teacher Coefficients That Incorporate Differing Information about Previous Grades

	Random effect		Fixed effect	
Variable	**coef**	**t-value**	**coef**	**t-value**
Average P/T during school years	-0.015	-2.41	-.014	-1.16
Average P/T in grades 1-4	-.020	-2.69	-.026	-2.60
P/T in year of the test	-.008	-1.32	.014	1.57

ables are shown in table 1 for random and fixed effect models. The results show that including current year pupil/teacher ratio instead of information from previous years causes the coefficients generally to weaken in both random and fixed effect models and to change signs in one model.

The Investments That Do Matter

The long debate about the role of resources in education has finally shifted from whether money does matter to what kinds of investments do matter for what kinds of children. The earlier conclusions drawn from reviews of the nonexperimental literature (Hanushek 1994)—that money has not mattered due to the inefficiency of our public school system and its lack of incentives—appear flawed. Over the last 25 years, money invested in schools for regular education students has gone mainly to develop programs targeted at minority and disadvantaged youth, lower pupil/teacher ratios, and raise average teacher salaries. Evidence is emerging that at least two of these investments have paid off for minority and disadvantaged students—lowering pupil/teacher ratios and targeting resources to minority and disadvantaged children. However, at least part of the money used to reduce pupil/teacher ratio for students from families with higher SES levels—the majority of students—may have been spent inefficiently.

Still, the broad-ranging conclusions that money does not matter in education without substantial changes in the existing structure of and incentives in public education are contradicted by experimental evidence and the results presented here. Moreover, the evidence supporting these conclusions now appears to be based on poor model specifications. This leaves the more viable hypothesis—that money *does* matter if invested in the right programs and targeted toward minority and disadvantaged students (Grissmer, Flanagan, and Williamson 1998b; Grissmer 1999).

Implications for Future Methodology and Data Collection

We suggest several specific ways that data and methodology might be improved, as follows: building micro-level models of educational processes, conducting more experiments, and improving NAEP data. For the latter, we discuss such measures as using school district samples rather than school samples, collecting additional family variables, improving children's responses (especially with regard to reporting levels of parental education), collecting additional information from teachers, using state Census data to improve the individual level variables, using supplementary data from the Census, and collecting additional parent information.

Building Micro-level Models of Educational Processes

The results of either experimental or nonexperimental analysis are meant to provide the material for developing theories of educational processes and student learning that gradually incorporate wider phenomena in their purview. Eventually, these theories should accurately predict the results of empirical work and be able to make new predictions to guide future empirical work. Theories by their very nature are more robust than any set of experimental or nonexperimental studies since they incorporate results of multiple measurements and incorporate research across levels of aggregation. However, little theory building has been done in education.

Hierarchies exist in science whereby certain areas of science are derived from and built upon the knowledge in more basic science. For example, the science of chemistry relies partly upon basic knowledge in physics for explanation. The science of biology is partly built from knowledge of chemistry; and, within biology, molecular biology provides some basis for the applied science of medicine. Typically the ordering of these hierarchies is derived from the size of the basic building blocks studied. Physics studies elementary particles and atoms. Chemistry studies combinations of atoms. Biology studies complex combinations of atoms with certain structures (genes, etc).

Education is far up in the hierarchies of social science. It rests upon knowledge derived from psychology, cognitive and brain science, genetics, sociology, child development, psychopathology, and economics. It is one of the more complex "sciences" that depends on good basic science in the lower hierarchies. Without linking the knowledge from these more basic sciences, educational research will never have a solid foundation. Educational research needs to incorporate the findings of these more basic sciences in building its theories, data collections, and methodologies. An example is the need to understand why smaller class sizes seem to produce higher levels of student achievement and why the results are multiyear and can be either short- or long-term.

Research directed toward measuring class size effects has generally treated the classroom as a black box in which only inputs and outputs are needed and in which knowledge of the transforming processes inside are unimportant for purposes of measurement. The current analytical methods also isolate the cause and effects of class size reductions within precise time periods in a way that seems at odds with the more continuous, cumulative, and often delayed effects that occur in children's cognitive development. Reconciling the differences in experimental and nonexperimental evidence will probably require a far better understanding of the underlying mechanisms occurring in classrooms and the developmental process in students that determine achievement.

In the case of class size, we need a theory of classroom and home behavior of teachers, students, and parents that answers why smaller classes might produce higher achievement in both the short and the long term. Initially we need to understand what teachers and students do differently in large and small classes and then whether these differences can be related to the size of short-term achievement. Perhaps the more difficult area of theory will be to explain gains long after the end of an intervention. An early intervention either has to change cognitive, psychological, or social development in important ways or change the future environment (e.g., peers, families) that affects the individual. Possibilities range from changes in brain development to learning different ways of interaction with teachers and peers to developing different study habits to being in different peer groups years later.

Answering these types of questions not only requires different types of data collection, but also requires understanding much about psychology, child development, and individual behavior (teacher and student). We provide some simple examples in the appendix at the end of this paper of the types of modeling and data collection that spring from alternate hypotheses about why smaller class sizes work.

One type of theory-building would use *time on task* as a central organizing concept in learning. A secondary concept involves the productivity and optimal division of that time among the different alternatives: new material through lectures, supervised and unsupervised practice, periodic repetition, and review and testing.[19] Students have a wide variance in the ways they spend time in school and at home, and it is likely that home time can substitute for specific types of teacher time.

Some research suggests that significant differences may exist in the amount of instructional time and the ways in which it gets used across different types of classes and different teachers and by students with different characteristics (Molnar et al. 1999; Betts and Shkolnik 1999a; Rice 1999). A theory of learning needs to be developed that incorporates school and home time and the various tradeoffs and differences that exist across teachers, classrooms, and SES levels. Such a theory would generate a number of testable hypotheses for research, which would then allow better and probably more complex theories to be developed. Such theories would then provide guidance as to what research is important to undertake.

Such theory-building would mandate linking several disparate and isolated fields of research in education. There is micro-research involving time on task, repetition, and review in learning specific tasks. There is research on teachers in classrooms. There is research on homework and tutoring. There is research on specific reading and math instructional techniques. There is research on class size and teacher characteristics. Theorists can begin to understand these disparate areas and suggest theories that can explain the empirical work across these areas. Such linkages seem essential to future progress.

Finally, cognitive development may have patterns of development similar to other areas of development in children, since brain development seems to be central to each type of development. There is much research on patterns of physical, emotional, and social development in children from birth, 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 oc-

[19] This approach is best exemplified in Betts and Shkolnik (1999a, 1999b) and Betts (1997).

cur even in highly risky situations. Much can be learned from this literature that can help prevent researchers from making poor modeling assumptions.

Need for Experiments

A major question raised by many other researchers, and currently under discussion, is the role of experimentation in educational research and other areas of social science (Burtless 1993; Boruch and Foley 1998; Boruch 1997; Hanushek 1994; Heckman and Smith 1995; Ladd 1996a; Jencks and Phillips 1998). Many interesting and complex issues arise in thinking about future experimentation, but consensus is emerging on the need for *more* experimentation in education.

Certainly, the value of the Tennessee experiment suggests that a selected number of social experiments may considerably add to our consensus knowledge in education. Besides the accuracy of the direct results, experiments tell us how to get more reliable results from nonexperimental data. Although expensive to carry out, experiments may be cheap compared to the costs of ineffective educational policies.

However, experimentation is much easier in smaller settings than in the classic, large-scale social experiments such as that produced in Tennessee. A very simple set of experiments could be designed around classroom- and school-level variables that would be much easier to carry out, yet could provide a better underlying base of information on which to build educational theories. For instance, simple experiments that divide children who miss a particular test question into two remediation groups with retesting could help locate the cause of missed questions and help develop efficient methods of remediation.

Improving the NAEP Data

The NAEP data are becoming so central to issues in both educational and social policy that priority should be given to significant expansion and improvement. We address two issues with respect to the NAEP data: (1) redesign of the sample to be district- rather than school-based and (2) improving family variables.

A School District NAEP Sample

The hypothesis suggested here implies that the lack of historical data on schooling variables may prove to be a barrier to unbiased results with individual level NAEP data. Here we focus on one option that would improve the aggregate data analysis possible with NAEP data. If NAEP could become a school district sample rather than a school sample, then historical data from school districts (not available at the school level of aggregation) could be used in the formulation of variables.

A district level sample would also result in improved family variables in NAEP data, since Census data would be available for most school districts. Currently, family variables in NAEP cannot be improved with Census data at the school level because privacy concerns prohibit their use within school areas. A school district sample would also address another NAEP deficiency—namely, the absence of several educational policy variables not available at the school level, such as per pupil spending. A much wider and better defined set of educational policy variables is readily available at the school district level and is already collected. Thus, a school district, rather than school level, NAEP sample would be desirable from the standpoint of improving family controls and educational policy variables.

A straightforward random sample of students at the district level would involve additional administrative costs, because the districtwide student universe would be needed and administration of tests would have to occur across many schools or involve assembling students from many schools in a central location. Such a sample would also have the disadvantage that, while Census and educational policy data would be available at the district level, certain school level characteristics obtained from student data at the school level would be missing. For instance, the school level sample of students is often used to define the characteristics of peers and their families. So a trade-off would occur with a district sample in that the educational and family characteristics would improve, but less would be known about some of the local, school level characteristics. Much of this missing school level data could probably be collected using enrollment data available at the school level. For instance, instead of using the sample of 20 students per school to estimate percentage minority, this figure would be obtained from schoolwide enrollment data.

Another change that would occur with a district sample would be that the sample of teachers surveyed would increase substantially. Currently, a typical classroom sample is 10–25 students, and a single teacher survey is collected. In a district sample, there would be few students selected from the same classroom, so the teacher sample would approach more closely the size of the student sample. The larger teacher sample would have some advantages besides increased size. The desired teacher variables are the characteristics of all teachers of the students from the time they entered school. The current teacher sample of one to two teachers per school, since it is a very small sample, is a very weak proxy for the characteristics of teachers at the school or the characteristics of all previous teachers of the students. Obtaining a much larger sample of teachers at the district level would provide a better proxy for the kinds of teachers likely to have taught in the district.

It may be possible to combine school level and district sampling to obtain a reasonable sample for each. About one-half of public school districts have fewer than 1,000 students and only one or two elementary schools per district. Thus, this sample of school districts would be close to the size of a school sample. However, these districts constitute only about 6 percent of total students. At the other end of the spectrum, there are about 300 districts with over 20,000 students, which account for nearly one-third of all students. In these districts, the number of schools ranges from about 30 to over 600. In most of these districts, a district sample could be drawn based on samples of schools, with 5–10 students per school. The remaining 60 percent of students are in school districts where some limited clustering by school could occur, but a sound district sample would probably have to include students from most schools.

However, it may be feasible to design a joint district- and school-based sample that samples fewer students per school. Such a sample would have several analytical advantages. It would contain an additional hierarchy in the sample—the district level, where extensive and better data exist on families and schools. It could still contain school-based samples, but with fewer students per school. It would also enlarge the number of teachers surveyed. Such a sample design would, however, entail additional costs since more schools would be sampled, district samples would require more effort at developing universe files, and more teachers would be surveyed.

The question is whether the analytical advantage would be worth the additional cost. To answer this question, we suggest a two-stage feasibility analysis in which a preliminary assessment by a group of statisticians and researchers would be performed to see whether serious barriers exist, to develop preliminary cost estimates, and to better define the analytical advantage. This group would either recommend a more detailed study and assessment or make the judgment that the analytical advantage is probably not worth the cost.

One of the chief advantages of moving to a district sample is that comparisons of scores could be made for major urban and suburban area school districts. It is the urban school systems that pose the largest 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 sample sizes would be much larger than at the state level and could be expected to provide more reliable results than for states.

Improving Family-level Variables

The primary objective of NAEP has always been seen as monitoring trends in achievement rather than explaining those trends. One result of this philosophy is that few family variables have been collected with NAEP. Compared with family data collected with other national achievement data or on other government surveys dealing with children's issues, NAEP collects very few family variables. In addition, the quality of the family variables collected has always been questioned since they are reported by the students tested. The perception of weak family variables may partially explain why NAEP scores have not been utilized more frequently in research on educational and social policies.

We have compared the accuracy of NAEP family data with Census data at the state level and analyzed the sensitivity of our estimates with state NAEP data with NAEP variables, Census variables, and SES variables formulated from parent-reported NELS data (Grissmer et al. 1998). Not surprisingly, we find that NAEP variables for race and family type (single-parent or two-parent) match Census data well, once differences in the samples are accounted for. However, students substantially inflate their parents' education level at the college level. Fourth graders report 58 percent of their families include a college graduate compared to 26 reported in the Census; comparable figures for eighth graders are 42 percent compared to 25 percent in the Census. However, reports of "high school only" and "not a high school graduate" are much more accurate. Students appear to be unable to distinguish between "some college" and "college graduate"—and individuals using NAEP data should combine these two categories when using the data.

There are several ways that the family variables can be improved in the NAEP data collection. We describe six increasingly complex options.

Collecting additional variables from children. There are two variables that are strongly significant in equations linking family characteristics and achievement that can be easily included and that children probably could report with some accuracy. The first variable is family size (number of siblings), which should present little problem for student reporting. The second is current age of mother. The age of mother combined with the child's age would enable the variable of age of mother at birth to be computed. Some pretesting may be required to determine the method of asking these questions, but even reporting mother's age in gross categories—five-year groupings—would be an improvement.

Recent research is finding that two-parent families with a stepparent do not have similar effects as do two biological parents (McLanahan and Sandefur 1994). The effects on children from a family including a stepparent appear to be closer to single-parent effects than to living with two biological parents. So information that could distinguish two-parent biological families from those with a stepparent would be useful. Adding a question on whether the parents are divorced is one approach. Asking separate questions about living with each parent is another approach.

One other variable that should be considered is locus of control. Locus of control is derived from a set of questions focusing on the perceived ability to affect life events. There are now more specific sets of questions that focus on specific events or conditions such as school performance. Locus of control has been collected in the NELS and NLSY data sets and is strongly statistically significant in equations relating achievement to family characteristics after all the common family characteristics are entered.

Improving children's responses. It appears that students have the least knowledge about post–high school education levels of parents. One hypothesis is that children have simply never asked parents about education level. Another is that parents report inaccurate levels of education to children, somewhat inflating their own level of education. In the former case, it may be possible to have children formally or informally ask parents prior to the test. This could take the form of a simple request before the test or a more formal written form for the parents to fill out. Pretesting this approach could help determine which hypothesis is causing the inaccuracy in reporting.

Collecting supplemental data from teachers. While individual level parental characteristics are desirable, teachers of NAEP students currently fill out an extensive survey that could be used to obtain family information. Teachers currently do not provide information concerning the socioeconomic characteristics of their students. Teachers could be asked several questions concerning the characteristics of the groups of students in their classes that might improve the data on family characteristics. These questions would take the form of identifying percentages of the students that fall into various categories. Income levels would probably be the most useful information. Giving teachers broad categories of income could prove better than the category of free and reduced price lunch as a control for family income. Items could include estimates for nearly all the important family variables. Such information could be first collected on a trial basis at low additional cost, perhaps for one year and utilized to see whether it improves the models.

Using state census data to improve individual level variables. We have utilized Census data to improve NAEP family variables at the state level. If NAEP data were only to be analyzed at the state level, the Census data combined with NAEP data could probably provide good estimates of all family background variables. However, the real value of NAEP data lies in the individual level data, and direct Census data have not been available at that level. So similar techniques cannot be used to directly derive school or individual level Census estimates.

It is possible to improve some of the reported NAEP variables at the school and individual levels by using the knowledge gained from state level comparisons. State level comparisons provide information about the accuracy of items such as parental education, and this information can in a limited way be used to impute better estimates to individual level variables. One simple application of this is to combine high school plus and college as a single category.

Further regressions across states linking the NAEP and Census estimates can provide information about how differences are connected to other family characteristics. For instance, the errors in reporting family education may be greater in states with high minority populations and lower incomes. This kind of information may be useful to impute better values at the individual level data. Such work would seek to better identify the types of students who report accurate and inaccurate data. However, while this approach should be tried, it would probably not result in dramatic improvements in the quality of individual level data.

Using supplementary data collection. The key information about family characteristics at the school level that would improve NAEP data might also be gathered directly from Census data. While privacy concerns limit the data available from Census at individual levels, the U.S. Department of Education would probably be able to obtain from Census data the school level population characteristics, if school boundaries were available. This would have to be done in conjunction with the NAEP data collection by collecting school boundary data on maps. Many school districts may be sufficiently large to allow Census to provide school data aggregated from the block level. This option should certainly be explored with the Census Bureau, and its cost assessed. There are many commercial vendors who can provide such data if given maps for specific disaggregated areas. The relative cost of this option compared to the cost of NAEP would be low.

The Census data could provide almost all the important background characteristics at the school level. But it would only be for all families in the area—not just the characteristics of families with fourth graders, for example. But the data would be highly correlated. Such data also could not track well the changes over time. Finally, the data would also be biased to the extent that the student population is not defined by specific geographical boundaries. But the advantages of this method would be the relatively low cost and the ability to provide a much richer set of characteristics at the school level.

Limiting parental data collection. Parental data collection for NAEP has always been a politically controversial issue, so extensive data collection similar to the type of collection performed on other U.S. Department of Education surveys is probably not feasible. The NELS, for instance, collects data from parents in an extensive survey. We consider here the minimum level of information which parents could provide that would enhance the NAEP data. The primary reason for parental data collection is to strengthen the individual level data in NAEP. A simple one-page form with no more than five items could solve the major problem with NAEP family data. It would take no more than a minute or two to fill out. It would ask for the key family background variables necessary for achievement score equations that are not accurately provided by the student. They include education level of each parent, family income in categories, and age of each parent. While a more extensive survey could certainly provide useful information, this minimum level of information would allow considerably more confidence in the use of individual level NAEP data without placing an undue burden on parents or children.

Summary

The interdisciplinary nature and the inherent complexity of educational research contribute their own set of challenges, but an additional reason for the lack of success in building consensus in educational research is the low investment in educational R&D and more broadly on R&D on children. On average, the nation spends approximately 2–3 percent of its gross domestic product for R&D. However, this proportion is not uniform across sectors of the economy, but can vary from less than 1 percent to approximately 20 percent (pharmaceuticals and integrated circuits) (Grissmer 1996). Currently, we spend less than 0.3 percent of educational expenditures for R&D, and less than 0.3 percent of expenditures for children are directed toward R&D on children (Consortium on Productivity in the Schools 1995, Office of Science and Technology Policy 1997). Compared to other sectors, this is a very low investment in R&D. Perhaps the reported problematical quality of educational R&D is partly due to the insufficiency of funding, when compared to its inherent complexity (Grissmer 1996; Wilson and Davis 1994; Atkinson and Jackson 1992; Saranson 1990). Alternately, the low funding level might reflect the poor quality of R&D.

Successful R&D is the engine that drives productivity improvement in every sector of our economy. Thus, strong R&D in education is a prerequisite to continual improvement in our education system and in our children's well–being. Without solid R&D, we will continue to go through wave after wave of reform without clearly separating the successful from the unsuccessful.[20] It is difficult to see how American K–12 education can become world class unless our educational R&D begins to build a more solid foundation of knowledge concerning education. If R&D can begin to play the role that it does in virtually every other sector of our economy, then continual educational improvement can be taken for granted, just as continual improvement in automobiles, computers, and life expectancy is now taken for granted.

[20] It is not that some reforms may not have been effective or had an impact on educational outcomes. The history of student achievement and educational outcomes suggests that scores have risen over long periods of time—and that students of a given era always seem to outscore their peers of earlier eras (Neisser 1998; Rothstein 1998). Rather, R&D could considerably improve the efficiency of the process of sorting the various reform initiatives and ensuring that the best are saved and the worst discarded.

Appendix

Simple Process Models of Class Size Effects

We start here by developing some simple models of the mechanism within classrooms that might cause class size effects and follow the implications of these assumptions on how we should specify models and why class size effects might be expected to have fairly wide variance. We do this simply to show that an important link is missing, a link that can guide us in specifying models and interpreting results of previous studies. If class size effects are produced by the kind of mechanisms assumed here, it implies that actual class size effects should have a wide variance and that some of the model specifications that were thought to be best actually can provide highly biased results.

Reductions in class size must change processes that occur in the classroom in order to have impacts on achievement. These differences in process that occur within smaller classes appear to determine whether class size affects achievement at all, whether effects are large or small, and whether effects widen or stay constant over several grades (Murnane and Levy 1996). In addition, the design of assessment instruments can determine whether class size effects are present in measurements.

Unless we know what processes change and how achievement is assessed, we cannot determine what model specifications and estimation techniques are appropriate. Since the data to determine what processes change in smaller class sizes are generally not collected, it will be difficult to sort out the reasons for the wide variance in the previous literature. We will discuss some simple, but extreme models to illustrate the point.

Demand for Teacher Individual Time

If we assume a "college professor" lecture model of classroom procedure, where there is essentially little or no interaction between teacher and student either during or after class and administrative time is borne by teaching assistants, then class size makes no difference. In this case, there is no cost to the teacher in having more students in the class. Class size makes a difference only when we assume that some teacher time is taken up by individual students—either through questions, special academic assistance, disciplinary actions, or administrative time (grading homework). In this case, additional students add to the teaching workload. If teachers have a fixed amount of time, then adding students can result in less time for presenting material or less time for student assistance. Thus, the size of the class size effect should depend on the portion of time teachers spend dealing with individual students (in one way or another) vs. time spent in general, lecture style instruction. In general, the more students need individual time, the larger will be the class size effect—other things being equal.

A second consideration is the variance among different types of students in requiring individual attention. A reasonable assumption here is that higher ability students or those with higher levels of family resources (broadly defined)—on average—will require less individualized attention. Essentially, substitution is occurring between family resources and school resources. In families with more resources, more of the students' academic and psychological needs are addressed at home, requiring less attention at school. This can include simple things such as helping with homework, enhancing learning opportunities, tutoring, and addressing the child's behavioral problems. For lower ability children or those with fewer family resources, more individualized attention will probably need to occur at school in order for them to achieve learning.

Thus, one would expect that class size effects would be larger for classes with lower ability students or students with fewer family resources. This also implies that there will be maximum class size levels (thresholds) that allow all the productive individual attention required, and above which no further class size effects will occur. But this threshold will vary by level of family resources.

Teacher and Curriculum Decisions

A third consideration is the teacher's reaction to scarcity of time. Teachers continually make choices about how fast to proceed with the scheduled curriculum, how much time to allow for slower students vs. faster students, and how much time to put into individual instruction vs. lecturing. With more students per class, these decisions become critical in determining whether class size effects occur. One scenario is that teachers slow down the pace of instruction in response to time scarcity. Individualized instruction is maintained for slower students, but less material is covered for all students. So the net effect is to cover less material for the school year for the whole class. Here, one might expect to see class size effects for the higher ability students (less material covered), but less so for those of lower ability.

Another teacher strategy is to cover all the material throughout the year (more time lecturing) and spend less time in individual instruction with slower children. Here, average class scores should shift downward in larger class size, but in a different pattern. Scores of higher scoring students would not be affected, but lower scoring students would have lower scores.

A crucial consideration in measurement is how the curriculum is adjusted the next year in response to these teaching strategies. If the effect of larger classes is failure to cover all the material for all students in the class, the next year's curriculum may or may not include all the material. If the curriculum accommodates this and starts where the previous year left off *for each student*, then over many years there will be an increasing gap between children in larger and smaller classes, i.e., the size of the effect will depend on the cumulative years in smaller class size. Thus, if smaller classes were instituted in grades K–8, one would expect to see a widening gap with each grade.

On the other hand, the start of next year's curriculum could begin *uniformly for all students* regardless of the amount of material covered last year. It could be started at the point where the larger or smaller class sizes left off. If it starts where the larger class sizes left off, then the gain from extra material covered in the smaller class in the previous year is lost. Thus, no cumulative effect is present, but a uniform score difference will be present each year. Essentially the smaller classes will cover additional material each year, but the gain from the previous year will be lost.

If the curriculum for all students is set where the smaller class sizes left off—leaving a permanent gap in coverage for those in larger class sizes—then whether the effect is cumulative depends on the extent to which mastery of the previous grade's material is required to perform well in the current grade. For subjects like math and reading, earlier mastery is probably more essential, and a widening gap would occur over several grades, i.e., the annual gap in material coverage would cumulate, causing further deterioration in later scores. On the other hand, in subjects like history or geography, where earlier mastery may be less important, the previous year's gap plays no role in next year's score, and a constant class size effect would be expected by grade.

Design of Assessment Instruments

Another consideration affecting the size of the measured class size is how assessment tests are designed. Designers of norm-referenced assessment tests first sample students' current knowledge at a given grade and develop a battery of questions that attempt to span the entire domain. A set of questions is chosen that provides a continuous range of question difficulty such that a different percentage of children answers each question correctly. Some questions nearly all students answer, while some are included that only a small percentage answer.

However, the domain of knowledge can depend on the size of classes attended by students. It is possible in some circumstances to have extra material covered by smaller class sizes included in assessments, while in other circumstances the extra material will not be part of the test. For instance, if tests were developed five years ago based on the then-existent domain of knowledge, and class sizes have declined since that time resulting in more material covered, the assessment instrument may not pick up the class size effect. Similarly, if assessment instruments are designed with students in larger classes prior to experimentation with smaller class sizes, then it is possible for the effects of class size to be attenuated if the instruments do not reflect possible additional material covered by smaller classes. In general, instruments designed to measure students' knowledge across several grades, rather than within each grade, and "re-normed" more frequently will be less vulnerable to these kinds of design effects.

Some Implications for Measurement and Specification

The above discussion illustrates that the size of the effect, its measurement, and its interpretation can depend on what occurs differently within the classroom when larger and smaller class sizes occur and on how assessments are designed. It implies that actual effects could vary considerably depending on different levels of student demand for individual time, teacher strategy, the coordination of the curriculum (e.g., year to year by class size), the different dependence by subject on previous knowledge, and assessment design. It would not be surprising in our decentralized educational system that smaller class sizes generate a wide variety of teacher and curriculum responses. Thus, ambiguity of results may not be surprising. Moreover, we may never be able to sort previous studies into groups with similar classroom process controls because the data along these kinds of dimensions were never collected for previous studies. So much of the work with previous data collections lacking these variables may have to be discounted.

The specifications of previous models have rarely taken explicit account of expected effect differences by family resource levels or tested for whether effects were constant or widened by grade. The latter consideration is critical to determining how models should be specified. For instance, if the conditions are present for a constant rather than an accelerating gap by grade, value-added models that control for previous years' test scores can show null effects of class size even though effects are present each year (Krueger 1999b). Effects would show up only in the first year in which class sizes were changed, but not in subsequent years. Such models would pick up only grade-by-grade acceleration in score changes. Here, simple cross-sectional models by grade without control for previous scores would show the total constant and cumulative effect to each grade.

The processes discussed above may or may not be the actual ones that exist in classes to produce class size effects. They simply point to the need to develop theories of the mechanisms underlying class size effects and to collect the data to test different theories. While a limited number of existing data sets might be able to start this process, it is difficult to see how definitive results are possible without more experimentation with more robust data collection on a much wider set of variables. Only by sorting this out can we be confident that models are specified correctly, estimation techniques are appropriate, and interpretations are accurate.

References

Advisory Group on the Scholastic Aptitude Test Score Decline. (1977). *On Further Examination.* New York: College Entrance Examination Board.

Achilles, C.M., Nye, B., Zaharias, J., and Fulton, B. (1993, January). *The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–91): A Legacy from Tennessee's Four-Year (K–3) Class Size Study (1985–90), Project STAR.* Paper presented at the North Carolina Association for Research in Education. Greensboro, North Carolina.

Armor, D. (1995). *Forced Justice: School Desegregation and the Law.* New York: Oxford University Press.

Atkinson, R., and Jackson, G. (Eds.). (1992). *Research and Education Reform: Roles for the Office of Educational Research and Improvement*. Committee on the Federal Role in Education Research, Washington, DC: National Academy Press.

Barnett, S. (1995, winter). Long-term Effects of Early Childhood Programs on Cognitive and School Outcomes, *The Future of Children 5*(3): 25–50.

Betts, J. R. (1997). *The Role of Homework in Improving School Quality,* Working manuscript. San Diego: University of California.

Betts, J. R., and Shkolnik, J. (1999a, summer). Estimated Effects of Class Size on Teacher Time Allocation in Middle and High School Math Classes, *Educational Evaluation and Policy Analysis 21*(2): 193–214.

Betts, J.R., and Shkolnik, J. (1999b). *The Effects of Class Size on Teacher Time Allocation and Student Achievement.* Working manuscript. San Diego: University of California.

Boruch, R. F. (1997). *Randomized Experiments for Planning and Evaluation.* Thousand Oaks, CA: Sage Publications.

Boruch, R. F., and Foley, E. (1998). The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials. In L. Bickman (Ed.), *Validity and Social Experimentation: Donald T. Cambell's Legacy.* Thousand Oaks, CA: Sage Publications.

Burtless, G. (1993). The Case for Social Experiments. In K. Jensen and P. K. Madsen (Eds.), *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*. Copenhagen: Ministry of Labour.

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

Cambell, J. R., Donahue, P. L., Reese, C. M., and Phillips, G. W. (1996). *NAEP 1994 Reading Report Card for the Nations and the States; Findings from the National Assessment of Educational Progress and Trial State Assessment* (NCES 95–045). Washington, DC: National Center for Education Statistics.

Cherlin, A. (1988). The Changing American Family and Public Policy. In A. Cherlin (Ed.), *The Changing American Family and Public Policy* (pp. 1–29). Washington, DC: The Urban Institute 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*. Washington, DC: U.S. Government Printing Office.

Consortium on Productivity in the Schools (1995). *Using What We Have to Get the Schools We Need*. New York: Teachers College Press, Columbia University.

Cook, M., and Evans, W. M. (1997). *Families or Schools? Explaining the Convergence in White and Black Academic Performance*. Working paper. University of Maryland.

Ferguson, R. F. (1998). Can Schools Narrow the Black-White Test Score Gap? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 318–374). Washington, DC: The Brookings Institution Press.

Ferguson, R. F., and Ladd, H. F. (1996). How and Why Money Matters: An Analysis of Alabama Schools. In H. F. Ladd (Ed.), *Holding Schools Accountable*. Washington, DC: The Brookings Institution Press.

Finn, J. D., and Achilles, C. M. (1990, fall). Answers and Questions about Class Size: A Statewide Experiment. *American Educational Research Journal 27*(3): 557–577.

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).

Flynn, J. (1987). Massive IQ Gains in 14 Nations: What IQ Tests Really Measure. *Psychological Bulletin 101*(2): 171–191.

Fuchs, V., and Rekliss, D. (1992). America's Children: Economic Perspectives and Policy Options. *Science 255:* 41–46.

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. (1996). *Education Productivity*. Washington, DC: Council for Educational Development and Research.

Grissmer, D. W. (1999, summer). Assessing the Evidence on Class Size: Policy Implications and Future Research Agenda. *Educational Evaluation and Policy Analysis 21*(2): 231–248.

Grissmer, D. W. (forthcoming). The Use and Misuse of SAT Scores. *Journal of Psychology, Law and Public Policy.*

Grissmer, D. W., Flanagan, A., and Kawata, J. (1998). *Assessing and Improving the Family Characteristics Collected With the National Assessment of Educational Progress*. Santa Monica, CA: RAND.

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., Flanagan, A., and Williamson, S. (1998a). 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: The Brookings Institution Press.

Grissmer, D. W., Flanagan, A., and Williamson, S. (1998b). Does Money Matter for Minority and Disadvantaged Students?: Assessing the New Empirical Evidence. In W. Fowler (Ed.), *Developments in School Finance: 1997* (NCES 98–212) (pp. 13–30). U.S. Department of Education, Washington, DC: U.S. Government Printing Office.

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., Kirby, S. N., Berends, M., and Williamson, S. (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* (pp. 251–286). Washington, DC: American Psychological Association.

Hanushek, E. A. (1989). The Impact of Differential Expenditures on School Performance. *Educational Researcher 18*(4): 45–51.

Hanushek, E. A. (1994). *Making Schools Work: Improving Performance and Controlling Costs*. Washington, DC: The Brookings Institution Press.

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

Hanushek, E. A. (1999, summer). Assessing the Empirical Evidence on Class Size Reductions from Tennessee and Nonexperimental Research. *Educational Evaluation and Policy Analysis, 21*(2): 143–163.

Hanushek, E. A., and Jorgenson, D. (1996). *Improving America's Schools: The Role of Incentives.* Washington, DC: National Academy Press.

Hanuskek, E. A., and Rivkin, S. G. (1997, winter). Understanding the Twentieth-Century Growth in U.S. School Spending. *Journal of Human Resources 32*(1): 35–67.

Hauser, R. M. (1998). Trends in Black-White Test Score Differentials: Uses and Misuses of NAEP/SAT Data. In U. Neisser (Ed.), *The Rising Curve: Long-term Changes in IQ and Related Measures* (pp. 219–250). Washington, DC: American Psychological Association.

Haveman, R., and Wolfe, B. (1994). *Succeeding Generations: On the Effects of Investments in Children*. New York: Russell Sage Foundation.

Haveman, R., and Wolfe, B. (1995, December). The Determinants of Children's Attainments: A Review of Methods and Findings. *Journal of Economic Literature 33*: 1829–1878.

Heckman, J. J., and Smith, J. (1995, spring). Assessing the Case for Social Experiments. *Journal of Economic Perspectives 9*(2): 85–110.

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 Greenwald, R. (1996). Have Times Changed? The Relation between School Resources and Student Performance. In G. Burtless (Ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (pp. 74–92). Washington, DC: The Brookings Institution Press.

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

Hedges, L. V., and Nowell, A. (1998). Black-White Test Score Convergence Since 1965. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 149–181). Washington, DC: The Brookings Institution Press.

Herrnstein, R. J., and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: Free Press.

Jencks, C. (1992). *Rethinking Social Policy: Race, Poverty, and the Underclass*. Cambridge, MA: Harvard University Press.

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

Jones, L. V. (1984, November). White-Black Achievement Differences: The Narrowing Gap. *American Psychologist 39:* 1207–1213.

Karoly, L. A., Greenwood, P. W., Everingham, S. S., Hoube, J., Kilburn, M. R., Rydell, C. P., Sanders, M., and Chiesa, J. (1998). *Investing In Our Children: What We Know and Don't Know About the Costs and Benefits of Early Childhood Interventions.* Santa Monica, CA: RAND.

Karweit, N. (1989). Effective Kindergarten Programs and Practices for Students at Risk. In R. Slavin, N. Karweit, and N. Madden (Eds.), *Effective Programs for Students At Risk* (pp.103–142). Boston: Allyn and Bacon.

Koretz, D. (1986). *Trends in Educational Achievement.* Washington, DC: Congressional Budget Office.

Koretz, D. (1987). *Educational Achievement, Explanations and Implications of Recent Trend*s. Washington, DC: Congressional Budget Office.

Krueger, A. B. (1998, March). Reassessing the View That American Schools Are Broken. *Economic Policy Review*: 29–43.

Krueger, A. B. (1999a). An Economist's View of Class Size Reductions. Unpublished paper. Princeton University. Princeton, NJ.

Krueger, A. B. (1999b, May). Experimental Estimates of Education Productions Functions. *Quarterly Journal of Economics 114:* 497–532.

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

Ladd, H. F. (1996b). Introduction. In H. F. Ladd (Ed.*), Holding Schools Accountable* (pp. 1–22). Washington, DC: The Brookings Institution Press.

Lankford, H., and Wyckoff, J. (1996). The Allocation of Resources to Special Education and Regular Instruction. In H. F. Ladd (Ed.), *Holding Schools Accountable* (pp. 221–257). Washington, DC: The Brookings Institution Press.

Linn, R. L., and Dunbar, S. (1990). The Nation's Report Card Goes Home: Good News and Bad News and Trends in Achievement, *Phi Delta Kappan 72*(2): 127–133.

Masten, A.S. (1994). Resilience in Individual Development: Successful Adaptation Despite Risk and Adversity. In M.C. Wang and E.W. Gordon (Eds.), *Educational Resilience in Inner City America: Challenges and Prospects*. Hillsdale: Lawrence Erlbaum Associates.

McLanahan, S., and Sandefur, G. (1994). *Growing Up with a Single Parent: What Hurts, What Helps*, Cambridge, MA: Harvard University Press.

Miller, K. E., Nelson, J. E., and Naifeh, M. (1995). *Cross-State Data Compendium for the NAEP 1994 Grade 4 Reading Assessment* (NCES 95–157). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., and Ehrle, K. (1999, summer). Evaluating a Pilot Program in Pupil/Teacher Reduction in Wisconsin. *Educational Evaluation and Policy Analysis 21*(2): 165–178.

Mosteller, F. (1995). The Tennessee Study of Class Size in the Early School Grades. *The Future of Children 5*(2): 113–127.

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: Data from the National and Trial State Assessments* (NCES 93–261). Washington, DC: National Center for Education Statistics.

Murnane, R. J., and Levy, F. (1996). Evidence from Fifteen Schools in Austin, Texas. In G. Burtless (Ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*. Washington, DC: The Brookings Institution Press.

National Research Council (1989). *A Common Destiny: Blacks and American Society*. Washington, DC: National Academy Press.

Neisser, U. (Ed.). (1998). *The Rising Curve: Long-term Changes in IQ and Related Measures*. Washington, DC: American Psychological Association.

Nye, B., Hedges, L., and Konstantopoulos, S. (forthcoming). *The Effects of Small Class Size on Academic Achievement: The Results of the Tennessee Class Size Experiment.*

Nye, B., Hedges, L., and Konstantopoulos, S. (1999, summer). The Long-term Effects of Smaller Class Size: A Five-year Follow-up of the Tennessee Experiment. *Educational Evaluation and Policy Analysis 21*(2): 127–142.

Office of Science and Technology Policy. (1997, April). *Investing in Our Future: A National Research Initiative for America's Children for the 21st Century.* Washington, DC: The White House.

Orfield, G., and Eaton, S. (1996). *Dismantling Desegregation: The Quiet Reversal of Brown v. Board of Education*. New York: The New Press.

Phillips, M., Crouse, J. and Ralph, J. (1998). Does the Black-White Test Score Gap Widen After Children Enter School? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 229–272). Washington, DC: The Brookings Institution Press.

Popenoe, D. (1993). American Family Decline, 1960-1990: A Review and Appraisal. *Journal of Marriage and the Family 55*: 27–555.

Raudenbush, S. W. (1994). Random Effects Models. In H. Cooper and L. V. Hedges, (Eds.), *The Handbook of Research Synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W., and Wilms, J. D. (1995, winter). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics 20*(4): 307–335.

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). Washington, DC: National Center for Education Statistics.

Rice, J. K. (1999, summer). Estimated Effects of Class Size on Teachers' Time Allocation in High School Mathematics and Science Courses. *Educational Evaluation and Policy Analysis 21*(2): 215–230.

Ritter, G., and Boruch, R. (1999, summer). The Political and Institutional Origins of the Tennessee Class Size Experiment. *Educational Evaluation and Policy Analysis 21*(2): 111–126.

Rock, D. A. (1987). The Score Decline from 1972–1980: What Went Wrong? *Youth and Society 18*(3): 239–254.

Rothstein, R. (1998). T*he Way We Were: The Myths and Realities of America's Student Achievement*. New York: The Century Foundation Press.

Rothstein, R., and Miles, K. H. (1995). *Where's the Money Gone? Changes in the Level and Composition of Education Spending*. Washington, DC: Economic Policy Institute.

Rutter, M. (Ed.). (1988). *The Power of Longitudinal Data.* Cambridge: Cambridge University Press.

Saranson, S. B. (1990). *The Predictable Failure of Educational Reform: Can We Change Course Before It's Too Late?* San Francisco, CA: Jossey-Bass Publishers.

Schofield, J. (1995). Review of Research on School Desegregation's Impact on Elementary and Secondary School Students. In J. A. Banks and C. A. McGee-Banks (Eds.), *Handbook of Research on Multicultural Education* (pp. 597–617). New York: McMillan Publishing.

Smith, M. S., and O'Day, J. (1991). Educational Equality: 1966 and Now. In D. Verstegen and J. Ward (Eds.), *Spheres of Justice in Education: The 1990 American Education Finance Association Yearbook* (pp. 53–100). New York: HarperCollins.

Stacey, J. (1993). Good Riddance to 'The Family': A Response to David Popenoe. *Journal of Marriage and Family 55*: 545–547.

United States Department of Agriculture (1997). *Estimates of the Costs of Raising Children*. Washington, DC: Author.

Wells, A., and Crain, R. (1994, winter). Perpetuation Theory and the Long-term Effects of School Desegregation. *Review of Educational Research 64*(4): 531–555.

Wilson, K. G., and Davis, B. (1994). *Redesigning Education*. New York: Henry Holt and Company.

Wilson, W. (1987). *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago, IL: The University of Chicago Press.

Word, E., Johnston, J., and Bain, H. (1994). *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985–1990*. Nashville: Tennessee Department of Education.

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.

Zill, N., and Rogers, C. (1988). Recent Trends in the Well-Being of Children in the United States and Their Implications for Public Policy. In A. Cherlin (Ed.), *The Changing American Family and Public Policy* (pp. 31–115). Washington, DC: The Urban Institute Press.

Response: Guidance for Future Directions in Improving the Use of NAEP Data

Sylvia T. Johnson[1]
Howard University

The issues of how to meaningfully use state National Assessment of Educational Progress (NAEP) data to assess and improve student achievement are the foci of the Grissmer and Flanagan and Raudenbush papers. They are exciting in their ramifications for future research and policy directions. The following discussion briefly describes NAEP and the whole idea of state-level data, then proceeds to review these papers in the context of their value in providing strategies for making these data more useful and informative assessments of national educational progress.

The assessment—as well as the improvement—of student achievement has long been a focus of educational policy at the state and the national levels and in the front lines of local school districts. With different emphases at different points in time, NAEP was originally designed as “the nation’s report card” to provide information on student achievement in subjects widely taught in public schools, for the nation as a whole and for specific demographic and geographic subgroups. However, in the first iterations of NAEP back in the 1970s, the regional subgroups were large, each including several states. It was not until the introduction of the Trial State Assessment (TSA) in 1990 that a sampling and administration structure was developed which allowed for the direct comparison of states with one another. Such between-states comparisons were not the explicit intent of the program. Rather, the TSA was intended to allow each state to compare its performance with that of the nation as a whole or perhaps with similar states in its own geographic region.

[1] The author is Professor, Research Methodology and Statistics at the School of Education, Howard University, and a principal investigator for the Center for Research on the Education of Children Placed at Risk (CRESPAR), an OERI-funded research center. She may be reached at Howard University, 2900 Van Ness Street NW, 116 Holy Cross Hall, Washington, DC 20008.

Actually, the involvement of states in NAEP is not new. An association of state educators, along with a core working group of the nation's top psychometrics scholars, were involved in the original conception and planning of NAEP.[2] They wanted to implement a national assessment program to document student progress in a manner which would not pose a threat to lower-performing district participation. To help ensure a low-key, relatively nonintrusive assessment program, NAEP results reported the percent of students who correctly answered each "exercise," as the assessment items were termed. Results were reported for geographic areas, and national samples of students were identified at ages 9, 13, and 17. NAEP currently assesses samples of students in grades 4, 8, and 11, but also reports trends for both age and grade for cross-sectional samples from about 1970 to the present. The initial trend year varies according to the time at which trend samples were introduced in each subject matter area (Campbell, Voelkl, and Donahue 1997).

In fact, the actual implementation of the Trial State Assessment has had a marked effect on how we measure student achievement. First, a motivational effect seems apparent in the TSA scores: they are a bit higher than regular NAEP scores. Second, certain states were anxious about their comparative standings; therefore, the "multiple comparison charts" show which unadjusted state mean differences were statistically significant from one another, as well as which differences were in the range of what would be expected simply due to chance. In these tables, no adjustments were made for student and family characteristics or for school resources and teacher background differences, although the importance of these factors certainly had been demonstrated in research studies carried out by NAEP, as well as in analyses of NAEP data in the literature. The Raudenbush and the Grissmer and Flanagan papers both addressed this problem of more meaningful use of state TSA data to assess student achievement, and both papers focus on the Trial State Assessment of NAEP. The data are based on eighth grade mathematics proficiency estimates from the Trial State Assessment.

Response to the Raudenbush Paper

In his paper, Raudenbush proposed a synthesis of state results by developing models that include correlates of student proficiency within and between states. He began with a "Conceptual Model for State-level Policy Effects on

[2] Personal communication from W. E. Coffman, 1975.

Student Achievement" (figure 1), which shows state government acting on student achievement primarily through its effects on school resources and home backgrounds. His analysis thus began by using NAEP data to assess the contribution of school resources; e.g., school and teacher quality, and students' home background to student achievement. The analysis was done for each of the 40 participating states, thus providing within-state home and school correlates of mathematics proficiency for eighth graders. He found a substantial correlation between socially disadvantaged or ethnic minority status, parental education, and access to the key resources available for learning, specifically course-taking opportunities, positive school climate, qualified teachers, and cognitively stimulating classrooms. He noted that these findings are similar to other findings in the literature. Carrying this analysis to another level, Raudenbush found considerable variation in the patterns of these correlations across the 41 states participating in TSA. These findings provided estimates for the direct effects of schools and teachers on student achievement while controlling for the correlation between self-reported student background, school factors, and teacher practices.

Raudenbush's analytic approach offers far more useful information to states than the conventional means from the NAEP TSA. By comparing states on resources and educational opportunities, this work enables the examination of possible changes in policy that are likely to positively influence student achievement. This analysis utilizing hierarchical linear models demonstrates that only a small amount of residual variance exists between states that is not related to school resources and family background. It should be noted here that there is a wide range in the proportion of within-school variability within states, which Raudenbush points out is also apparent across states. But the within-state variation is worth the attention of individual states; and there is some work in this area, for example, William Cooley's paper on Pennsylvania (Beckford and Cooley 1993), which examines schools with sizable numbers of African American students in which these students score at or above the state mean on achievement measures, and which also cites other relevant investigations into these questions.

Given the demonstrated importance of school resources to achievement, the second part of the Raudenbush paper presented an examination of state differences in school resources. This work explores two questions: "Does the distribution of school resources likely reinforce or counteract inequalities arising from home environment? Do states differ, not only in the provision of resources, but also in the equity with which they are distributed?" This examination thus focused not just on resource differences between states, but also on how equitably resources are distributed within states. These within-state resource differences were then examined, not only in terms of student demographic background, but also in the interaction between resource differences, race-ethnicity, and parental education. Raudenbush's analysis and his illustrative plots (figures 4, 5, 6, and 7) show that the probability of access to key resources is a function of these background factors. Further, he found sizable differences attributable to student background, education, and ethnicity; and these differences were associated with access to resources related to school success. For example, factors such as teacher quality and experience, school climate, whether or not a school offered algebra, and whether students were assigned to math teachers who majored in math and who emphasized reasoning in their classroom instruction—these were related to success, as well as to the variables used as predictors in Phase I of the Raudenbush work. These same variables, when examined across states, result in the ellipses (figures 8 and 9) that visually show the relative access to resources provided to African American students. For example, figure 8 shows that in South Carolina and Mississippi, having parents who are college educated offers only modest advantage to students, and the advantage is about the same for African Americans and youth of other backgrounds.

The meaning and the utility of Raudenbush's findings and this methodology for states and districts, and perhaps also for schools, are substantial. First, the absence of large statistical differences between adjusted means should in no way encourage states that they are to do little. The kind of action that would be prompted from the states was a concern of the National Assessment Governing Board when it decided, when plans were made for reporting the 1990 results, to report unadjusted means only. Second, the Raudenbush analyses clearly show the importance of access to school resources for student achievement. In order to move toward equity for all students, these findings demonstrate unequivocally that resource accessibility is a key factor.

In terms of the current "affirmative action" debate, these findings also have important implications for many states, such as Florida and California. First, can a state logically expect proportional representation by ethnicity on college entry characteristics such as test scores and course-taking patterns when it has systematically limited access to resources available to students? Given the demonstrated relation of resources to measured achievement reported in the Raudenbush paper, it would seem only logical that a state, in making college entrance decisions, should take into consideration the relative access it provides to certain resources in elementary and secondary school. Such a procedure would need, certainly, a well-specified model and additional research.

Though the many issues are still in flux, reactions to affirmative action are often far too simplistic in their conceptions of how the specific mechanisms work in practice. Where race has served as a factor in the allocation of skilled teachers and other education resources, whether by design or not, this condition raises the question of how this method of allocation has to be considered when the measurable results of such allocations are evaluated.

The author's broad recommendations should be noted here. If student progress and subject matter proficiency are examined on a year-to-year basis, the extent to which the educational system (or even the school) differentially provides resources that support student progress should also be examined. This is the *opportunity-to-learn* concept. The soundly based but creative methodology that is employed in the Raudenbush work offers strong promise for helping us better understand how student background and resource access are interrelated, as well as when such access factors have been modified. The latter could be examined by extending the analysis over time so that the progress of states in modifying resource access could be followed and appraised. The author suggested extending the collection of data by NAEP to measures of resources and development indicators from these data. Such a direction seems logical and important, but it stands in opposition to current plans to release results more quickly and to collect fewer data from students, teachers, and schools than has been done in the past.

How can states and schools use these findings? To begin with, they can collect comparable data at the district and school levels so that the internal allocation of resources is more completely documented. They can also modify teacher assignments to provide more equitable distribution of highly skilled teachers, although such a goal may entail the need for financial incentives, along with improving the facilities and working conditions for teachers in some schools, and other strategies. For teachers, the range in access to *everything*—from professional development to clean bathrooms—is very great across schools, even in the same or nearby districts. State officials may lobby for increased state support to remedy access problems: they could target selected schools and districts, using these findings as a basis for the request. They may investigate how parental education operates to increase access to resources and provide help to parent groups in lower-access schools to develop strategies to bring about change for the better in their schools. They can recognize the broad scope of the documented inequities and develop a broad-based strategy, sustained over the long term, to simultaneously and progressively change the inequities in resource allocation that are so widely spread among states, systems, and schools.

Response to the Grissmer and Flanagan Paper

In their paper, "Moving Educational Research toward Scientific Consensus," Grissmer and Flanagan assert the need to improve the consistency and accuracy of results in educational research so that a basic knowledge base in education can be built, one that can be accepted by a diverse research community as well as by educational practitioners and policymakers. The authors assert that improving nonexperimental data, along with the associated methodology, may not be enough to achieve consensus. Rather, they contend that experiments are needed and, further, that experiments should often employ models such that the size of a given year's effect can be viewed as dependent on the current year's and previous years' effects. These findings should then be used to build micro-theories of educational process.

Grissmer and Flanagan dealt with the issue of the relation between school resources and school achievement in two major examples. The first example which they cite is the rapid change in NAEP scores of black students, especially from 1970 to 1988 or 1990. In the case of black student progress in NAEP scores, Grissmer and Flanagan gave credit to compensatory and developmental programs and school desegregation activities, but they did not offer conjecture regarding the score declines that occurred starting from 1988 or1990.

The Tennessee class size study is well presented by Grissmer and Flanagan; and the posing of a simple process model for these effects is a useful and important addition to the literature. The detail presented to amplify and explain how teacher reactions to the scarcity or abundance of class time may interact with student characteristics is thoughtful. Readers are encouraged to take the time to examine that part of the paper carefully, as it provides further support for the need of extensive data collection, either from teacher questionnaires, interviews, or classroom observations.

It is good to see the Tennessee class size experiment get the kind of attention toward implementation that it has deserved for some years. This long-term experiment received relatively little attention in the policy arena until recently, but now seems to be getting wide notice. Certainly, the Tennessee class size study has reached the ear of the President—it is, indeed, a factor in the call for many new teachers across the nation. Interestingly, the Tennessee study results from collaboration between the historically black university, Tennessee State University, and the State Department of Education, with important guidance from a participant in this conference, Jeremy Finn. This well-designed study made it possible to study the effects of smaller classes over time. Thus, it is an excellent model for the kind of work that needs to be done to develop and test theories in education.

In addition to their suggestions for extending and improving NAEP data collections, Grissmer and Flanagan suggest improving NAEP by collecting family characteristics at the school level, possibly using Census Bureau data to augment NAEP. There are some states and other jurisdictions which have used Census data and other federal reporting information to improve their estimates for allocation of social and economic services. A major problem in this work has been the adequacy of geo-coding (the coding of addresses and other location information) in the files proposed for use to improve estimation. The problem is especially severe in sparsely populated areas; namely, rural communities, older urban industrial zones that have lost population with the closing of plants, and small towns. A National Research Council panel has been examining problems of estimation of poverty in small geographic areas, and the U.S. Department of Education, through the National Center for Education Statistics, is a sponsor of this work. Working in close collaboration with the Census Bureau, the panel has been operating for about 3 years, has published three interim reports, and is working to complete a final report (National Research Council 1999). Their findings should be useful in the improvement of parameter estimation for many forms of resource allocation.

Grissmer and Flanagan also suggested a school district sample rather than a school sample for NAEP and the use of a longitudinal cohort. A district sample, though more expensive to collect, might enable comparisons of scores for urban and suburban districts within metropolitan areas of similar size, though this level of reporting is disallowed under current NAEP authorization. Now, of course, prior to TSA, NAEP was a low-stakes testing program. Since comparisons are now easily made between states, it is worth considering whether we might facilitate between-district comparisons, if solid methodology that gives consideration to resource allocation is used to interpret those differences. Such a change would raise the stakes for State NAEP as well as for national NAEP. Given the effect of other high-stakes testing programs and the fundamental role of NAEP as the nation's report card, such changes should be carefully reviewed.

The authors support the use of longitudinal cohorts. A longitudinal study would involve identifying students or at least forming blocks at the school or district level, but the advantages would need to be weighed along with costs. NAEP currently examines trends by retaining a common core of test items which are administered to cross-sectional grade level groups.

Conclusion

Now let us consider what these papers, taken together, tell us. Both studies point out the importance of a school's climate and culture to discipline. In our work at CRESPAR, the Center for Research on the Education of Students Placed at Risk, an OERI-funded research center located at Howard University and Johns Hopkins University, we are guided by a talent development model recently articulated in an article by my colleague, Serge Madhere (1998; see also Boykin 1996). This model has a number of points, the most important one of which is that all children can learn, given adequate opportunity and that their backgrounds and culture have strengths that can be built on to motivate and encourage student learning. Learning is more a function of coherent instruction than of a child's social origin. Motivation begets greater learning, which begets greater motivation. Nurturing is the key to motivation, especially at difficult transition points.

Both of these papers offer creative methodological approaches to the use of state-level data for school improvement and for theory building. Both demonstrate the importance of resource allocation for student achievement and the interaction between race and resources, and both imply procedures for increasing proficiency among African American students. Both imply the need for more complex NAEP data at the level of the student, the family, the teacher, the school, and the state. This emphasis, however, runs counter to the current push to simplify and speed up the data collection and reporting process. More complex data require more complex consideration before developing conclusions.

Both studies show the need to better understand the why's and how's of improving student achievement. For example, how do teachers use time? When they have more time per student, what are the features of resources that make them effective in influencing achievement?

These are important directions for researchers and policymakers to consider. Fortunately, there is much in these papers to help guide that progress.

References

Beckford, I. A., and Cooley, W. W. (1993). *The Racial Achievement Gap in Pennsylvania*. Pennsylvania Educational Policy Studies Paper Number 18. ERIC No. EDD 389760.

Boykin, A. W. (1996, April). *A Talent Development Approach to School Reform: An Introduction to CRESPAR*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Campbell, J. R., Voelkl, K. E., and Donahue, P. L. (1997). *NAEP 1996 Trends in Academic Progress: Achievement of U.S. Students in Science, 1969 to 1996, Mathematics, 1973 to 1996, Reading, 1971 to 1996, Writing, 1984 to 1996*. (NCES 97–985). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Madhere, S. (1999). Cultural Diversity, Pedagogy, and Assessment Strategies. *Journal of Negro Education 67*(3): 280–295.

National Research Council. (1999). *Estimation of Poverty in Small Geographic Areas* (Third Interim Report). Washington, DC: National Academy of Science Press.

Section II.
Using Longitudinal Data to Assess Student Achievement

Understanding Ethnic Differences in Academic Achievement: Empirical Lessons from National Data

Meredith Phillips
University of California, Los Angeles

In 1966, James Coleman published results from the first national study to describe ethnic differences in academic achievement among children of various ages. Since that time, we have made considerable progress in survey design, cognitive assessment, and data analysis. Yet we have not made much progress in understanding when ethnic differences in academic achievement arise, how these differences change with age, or why such changes occur.[1] The purpose of this paper is to highlight several reasons why we have learned so little about these important issues over the past few decades. I begin by reviewing recent research on how the test score gap between African Americans and European Americans changes as children age. I then discuss several conceptual and methodological issues that have hindered our understanding of ethnic differences in academic achievement. I raise these issues in the hope that we will make more progress toward eliminating the test score gap during the next decade than we have during the last.[2]

[1] I use the term "ethnic" to refer to the major ethnic and racial groups in the United States (namely, African Americans, European Americans, Latinos, Asian Americans, and Native Americans). Whenever the samples are large enough, I also consider variation within these socially constructed categories (for example, differences between Mexican Americans and Puerto Rican Americans).

[2] I thank Robert Hauser, Larry Hedges, Christopher Jencks, Jeff Owings, and Michael Ross for their comments on an earlier draft. I did not make all the changes they suggested, however; and they are in no way responsible for my conclusions. Please direct all correspondence to Meredith Phillips, School of Public Policy and Social Research, UCLA, 3250 Public Policy Building, Los Angeles, CA 90095–1656 or phillips@sppsr.ucla.edu.

Does the Achievement Gap Change as Children Age?

My colleagues and I recently analyzed data from a number of national surveys in order to estimate how the achievement gap changes as children age (see Phillips, Crouse, and Ralph 1998). Answering this question can help us understand the potential causes of the gap. Suppose, for example, that the black-white gap did not widen at all after first grade, even among black and white children who began school with similar skills. If that were the case, we might conclude that families, communities, preschools, or kindergartens were mainly responsible for the gap. On the other hand, suppose that the black-white gap did widen between the first and the twelfth grades, even among children who started school with similar scores. If that were the case, we might conclude that schools were mainly responsible for the gap. As it turns out, the "truth" seems to fall somewhere between these extremes.

Cross-sectional Results

One way to describe age-related changes in the black-white gap is to estimate the size of the gap in as many surveys as possible and then combine these estimates. We have done this with the national surveys listed in table 1. Figure 1a arrays the black-white math gaps from these surveys by age. The lines around the estimates show their precision. We can also array these gaps by year of birth, which shows the historical trend in the black-white math gap (see figure 1b). Because the black-white gap narrowed during the 1970s and 1980s, however, we need to make sure that age-related changes in the gap are not confounded with historical changes. In order to disentangle the effects of age from the effects of history, we estimated a multivariate model that controlled for the historical trend while estimating the age-related trend.[3] Table 2 presents these results. It shows the following: the black-white math gap widens by about 0.18 standard deviations between the first and the twelfth grades; the reading gap stays relatively constant; the vocabulary gap widens by about 0.23 standard deviations.[4] A gap of one standard deviation on the math or verbal SAT is 100 points. Therefore, our cross-sectional results imply that the black-white math and vocabulary gaps widen by the equivalent of just under 2

[3] For details on the sample and analysis, see Phillips, Crouse, and Ralph (1998).

[4] To obtain these estimates, multiply the coefficients in the first row of table 2 by 12 years of school.

Table 1. Data Sets Used in Meta-analysis

Acronym	Name	Test Year(s)	Grades Tested
EEO	Equality of Educational Opportunity Study	1965	1,3,6,9,12
NLSY	National Longitudinal Survey of Youth	1980	10,11,12
HS&B	High School & Beyond	1980	10,12
LSAY	Longitudinal Study of American Youth	1987	7,10
CNLSY	Children of the National Longitudinal Survey of Youth	1992	Preschool, K, 1,2,3,4,5
NELS	National Education Longitudinal Study	1988, 1990, 1992	8,10,12
PROSPECTS	Prospects: The Congressionally-Mandated Study of Educational Growth and Opportunity	1991	1,3,7
NAEP	National Assessment of Educational Progress	1971–1996	4,8,11

Figure 1a.
Standardized Black-White Math Gaps, by Grade Level

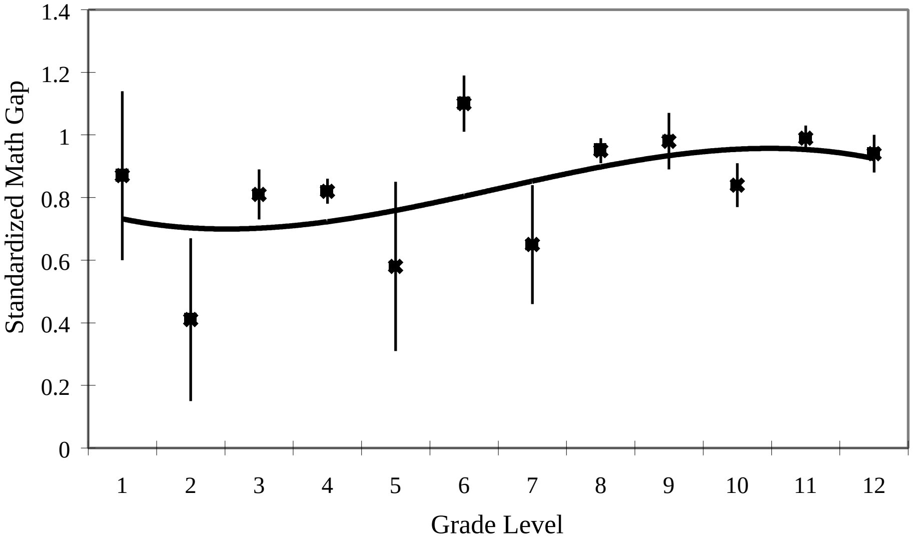


Figure 1b.
Standardized Black-White Math Gaps, by Year of Birth

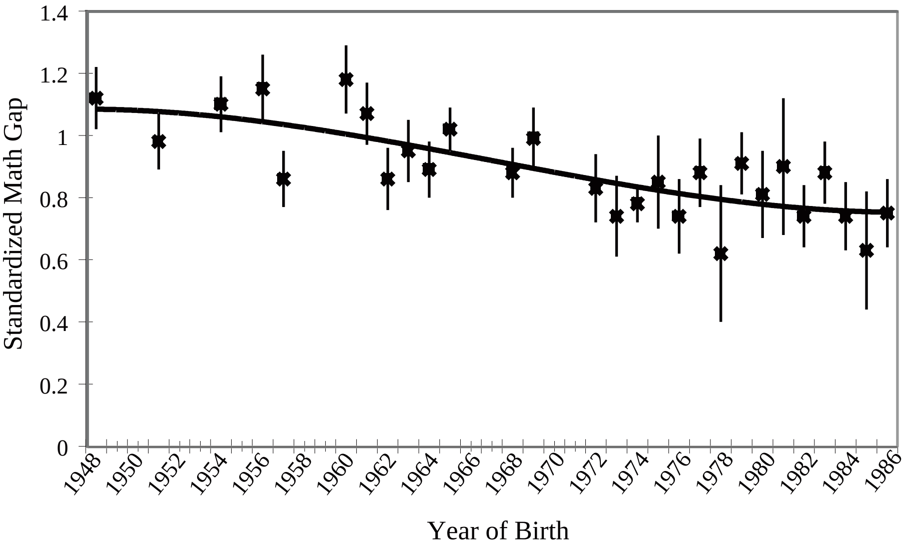


Table 2. Effects of Grade at Testing and Year of Birth on Black-White Test Score Gaps

		Dependent Variables					
		Mathematics (*N*=45)		Reading (*N*=45)		Vocabulary (*N*=20)	
Independent Variables		1	2	1	2	1	2
Grade level	B SE	.015 (.004)	...	.002 (.006)	...	.019 (.006)	...
Grades 1–6	B SE	...	.051 (.014)	...	-.011 (.023)	...	.034 (.012)
Grades 7–8	B SE	...	-.054* (.028)	...	.016 (.051)	...	.025 (.032)
Grades 9–12	B SE	...	.021* (.013)	...	.010 (.024)	...	-.018 (.017)
Month of testing	B SE	-.011 (.004)	-.007 (.004)	.003 (.005)	.000 (.007)	.015 (.018)	.011 (.018)
Year of birth before 1978	B SE	-.014 (.002)	-.014 (.002)	-.020 (.002)	-.020 (.002)	-.010 (.003)	-.011 (.003)
Year of birth after 1978	B SE	.002* (.006)	.004* (.005)	.020* (.009)	.018* (.010)	.031* (.011)	.039* (.012)

Table 2. Effects of Grade at Testing and Year of Birth on Black-White Test Score Gaps (continued)

		Dependent Variables					
		Mathematics (*N*=45)		Reading (*N*=45)		Vocabulary (*N*=20)	
Independent Variables		1	2	1	2	1	2
Longitudinal survey	B	-.039	-.043	-.069	-.063	-.346	-.273
	SE	(.033)	(.033)	(.047)	(.051)	(.157)	(.161)
IRT metric	B	.175	.149	.159	.174	.068	.000
	SE	(.033)	(.035)	(.046)	(.051)	(.082)	(.088)
Intercept	B	.765	.653	.746	.792	.889	.833
	SE	(.034)	(.054)	(.056)	(.092)	(.049)	(.057)
Adjusted R^2		.790	.815	.693	.680	.745	.806

NOTE: The dependent variables are standardized black-white gaps (i.e., $(\overline{W}\text{-}\overline{B})/SD_T$) computed from the surveys listed in table 1. The actual data appear in table 7A–1 in Phillips, Crouse, and Ralph (1998). Standard errors are in parentheses. The spline coefficients for grade level and year of birth show the actual slope for that spline. The spline standard error indicates whether the slope differs from zero. * indicates that the spline's slope differs significantly from a linear slope at the .05 level. Each gap is weighted by the inverse of its estimated sampling variance. See Phillips, Crouse, and Ralph (1998) for details on the other variables in this analysis. See pp. 118-19 of Pindyck and Rubinfeld (1991) for an introduction to spline (piecewise linear) models. See Cooper and Hedges (1994) for details on the meta-analytic methods used in this analysis.

SAT points a year, or by 18 to 23 SAT points over the course of elementary, middle, and high school.

These cross-sectional estimates have two advantages over longitudinal estimates. First, the data span nearly all grade levels, from early elementary school through late high school. No national longitudinal survey has ever tested children over an interval spanning both elementary school and high school. Second, because cross-sectional surveys do not follow students over time, they are less subject to attrition and thus tend to be more nationally representative than longitudinal surveys. A problem with our cross-sectional results, however, is that they combine data on children from different samples, who were assessed on different, possibly incomparable, tests. Another problem is that cross-sectional data cannot tell us whether the black-white gap widens among children who start school with the same skills. That question, which is central to the concern that schools may not be offering black and white students equal educational opportunities, can be answered only with longitudinal data.

Longitudinal Results

During the late 1980s and early 1990s, two national longitudinal surveys assessed students multiple times as they moved through school. The National Education Longitudinal Survey (NELS) is the more familiar of these studies. NELS is a large national survey that first tested eighth graders in 1988 and then retested them in 1990 and 1992. Prospects, a survey of two cohorts of elementary school students and one cohort of middle school students that began in 1991, is less familiar than NELS because it is not yet readily available to researchers. The Prospects data were collected mainly to evaluate the effectiveness of Chapter 1 (now Title I), but their secondary purpose was to describe yearly achievement growth during elementary and middle school. The youngest of the Prospects cohorts was first tested at the beginning of first grade and followed through the end of third grade. The middle Prospects cohort was first tested at the end of the third grade and followed through the end of sixth grade. The oldest cohort was tested at the end of seventh grade and followed through the end of ninth grade.

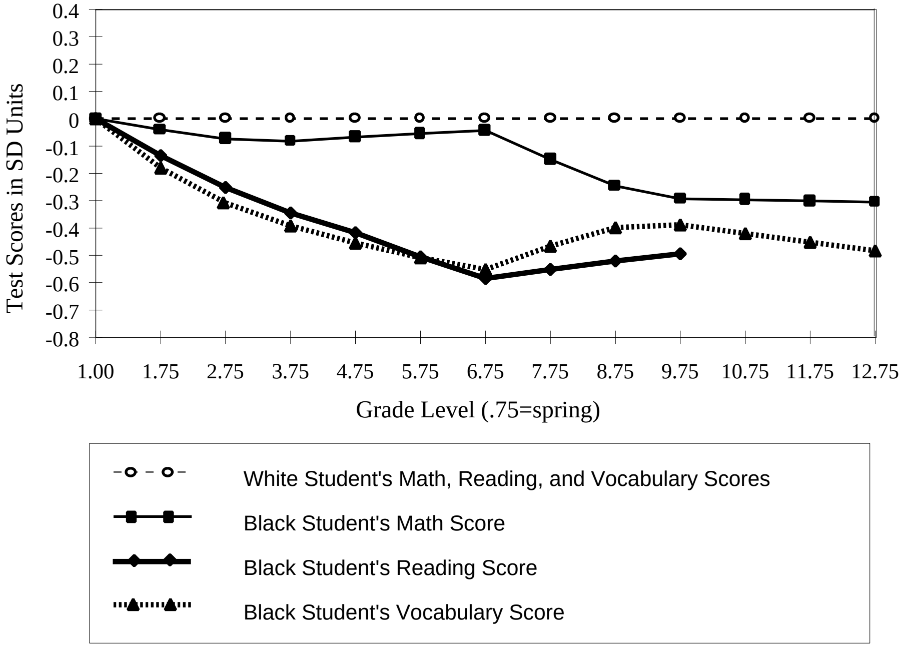
In order to understand achievement growth over an interval longer than four years, we have to piece together data from these different cohorts. My colleagues and I have used these data to estimate whether black children who start out with the same skills as whites learn less over the school years.[5] Our estimates are very imprecise because the Prospects sampling design was relatively inefficient and because we do not have data for every school year.[6] Nonetheless, our results suggest that African American children fall somewhat behind equally skilled white children, particularly in reading comprehension, and particularly during the elementary school years (see figure 2).[7] Taken together, we estimate that at least half of the black-white gap that exists at the end of twelfth grade can

[5] See Phillips, Crouse, and Ralph (1998) for details.

[6] Also, a very large percentage of the Prospects students left the study before the second and third waves. When Phillips, Crouse, and Ralph (1998) compared cross-sectional and longitudinal samples drawn from Prospects, however, they found that the mean black-white gap differed by less than 0.05 standard deviations across all tests. And although the longitudinal samples were more advantaged than the cross-sectional samples, *racial* differences in attrition were small and mostly involved regions of residence and urbanism. See chapter 3 of Phillips (1998) for more on nonrandom attrition in both Prospects and NELS.

[7] See Phillips, Crouse, and Ralph (1998) for a comparable figure using cross-sectional data, as well as for a figure that shows the imprecision of these predictions.

Figure 2. Predicted Test Scores for Two Students, One Black, One White, Who Both Started First Grade with True Math, Reading, and Vocabulary Scores at the Mean of the Population Distribution



be attributed to the gap that already existed at the beginning of first grade. The remainder of the gap seems to emerge during the school years.

This widening of the gap may not be attributable to schooling *per se*, however. Because of summer vacation, students spend only 180 days a year in school. Because neither Prospects nor NELS tested children in the fall and the spring of each school year, it is impossible to know how much of the gap that emerges over the course of schooling should be attributed to schools and how much should be attributed to summer vacations.[8]

In an ideal world, we would know precisely when ethnic differences in test scores first emerge and how they develop during the preschool years.

[8] Several other studies have examined summer learning patterns (e.g., Cooper et al. 1996; Entwisle and Alexander 1992, 1994; and Heyns 1978, 1987). Further, Prospects tested an unrepresentative subsample of students in the fall and spring of first and second grade. I review these results later in the paper.

We would also know how ethnic differences change both every school year and every summer. This information would help us identify the most important reasons why African American and Latino children score lower than whites and Asian Americans on math and reading achievement tests. Unfortunately, we are not close to knowing the answers to these seemingly basic, descriptive questions. In the remainder of this paper, I discuss several explanations for this knowledge gap.

Why Do We Know So Little?

The most obvious reason why we have made so little progress on the test score gap puzzle since 1966 is that most researchers have been reluctant to study it. Rather than directly tackling this politically sensitive subject, most scholars have tried to understand ethnic inequalities in academic skills by comparing socioeconomically disadvantaged students to advantaged students, by comparing students in high poverty schools to those in low poverty schools, or by comparing urban students to suburban students. All these comparisons pose interesting questions for social science. None, however, brings us closer to understanding ethnic differences in academic achievement because ethnicity does not overlap with social class and urbanism as much as most researchers assume.

Table 3 illustrates this problem. It shows the magnitude of the black-white test score gap among a national sample of eighth graders, according to the education and income levels of their parents, as well as the poverty and urbanism of their schools. If these other variables were adequate substitutes for race, the black-white gap would disappear after these variables were taken into account. The black-white gap does shrink, but it is still large *within* each of these categories. More sophisticated analyses that simultaneously control many family background variables yield similar results (see Phillips et al. 1998).[9] Racial and ethnic differences in test scores are *not* the same as SES differ-

[9] See also appendix C of Phillips, Crouse, and Ralph (1998) for data on how much the black-white gaps in Prospects and NELS shrink after controlling a number of common indicators of family background.

ences. Therefore, if we want to understand and eliminate racial and ethnic differences in test scores, we need to confront the problem directly.[10]

The results I presented on age-related changes in the black-white test score gap illustrate that even when we decide to focus explicitly on ethnic differences in academic skills, however, the available data are inadequate. In order to improve our understanding of the development and causes of ethnic differences in academic skills, supporters of educational surveys, such as the National Center for Education Statistics (NCES), need to do the following:

- Focus primarily on preschool and elementary school students rather than high school students;
- Assess at least a subsample of students in the fall and spring of *every* school year;
- Maximize measurement variation *within* each survey;
- Fund more than one survey of the same population at a time; and
- Remember that, because education begins before formal schooling, education surveys must also do the same.

I will elaborate on each of these points in turn.

The Importance of Early Schooling

Scholars who have studied ethnic differences in test scores have mostly focused on adolescents.[11] This is a mistake, for reasons that I will illustrate. It is, however, a reasonable mistake—at least among quantitative scholars—be-

[10] This does not mean, of course, that policies aimed at reducing the test score gap need be targeted at specific racial or ethnic groups rather than at low-scoring students in general. As Christopher Jencks and I argue in our introduction to *The Black-White Test Score Gap* (1998), the best policies seem to be those that help both blacks and whites, but help blacks more.

[11] See, for example, Ogbu (1978), Fordham and Ogbu (1986), Kao, Tienda, and Schneider (1996), Cook and Ludwig (1997), and Ainsworth-Darnell and Downey (1998). The main exceptions are Doris Entwisle and Karl Alexander, the founders of the Beginning School Study (BSS) in Baltimore, who continue to follow students who began first grade in 1982. Entwisle and Alexander's (1988, 1990, 1992, 1994) studies have been the source of most of our knowledge about the development and causes of black-white differences in academic achievement. Yet the BSS sample is relatively small, does not include enough schools to estimate between-school differences precisely, does not include Latinos or Asians, and may not generalize to other samples, because white students in the Baltimore public schools are not representative of white students nationally.

Table 3. Standardized Black-White Test Score Gap among Eighth Graders in NELS, by Parental Education, Income, School Poverty, and Urbanism

	Black-white gap on a combined math and reading test among eighth graders
In the overall population:	-0.80
Whose parents are:	
High school dropouts or graduates	-0.55
College graduates	-0.85
Whose parents' income is in the:	
Bottom fifth of the distribution	-0.53
Top fifth of the distribution	-0.53
Who attend schools in which:	
More than 40 percent of students are eligible for free or reduced lunch	-0.57
Fewer than 5 percent of students are eligible for free or reduced lunch	-0.65
Who attend:	
Urban schools	0.89
Suburban schools	-0.75

NOTE: The denominator of the gap is the weighted overall population standard deviation.

cause NCES has not, until very recently, supported surveys of elementary school students. NCES has conducted three large surveys of high school students: The National Longitudinal Study of 1972 (NLS-72), the High School and Beyond survey of 1980 (HSB), and NELS. All the data and documentation from these three studies are available to researchers who agree to abide by specific security provisions. Although the U.S. Department of Education has also supported two longitudinal surveys of elementary school students—the Sustaining Effects Study of 1976 and the Prospects study—the main purpose of these elementary school surveys was to evaluate Chapter 1, not to study learning during elementary school. Neither data set is widely available to the research community, and the quality of the data and documentation is much lower than in the high school surveys.[12]

[12] The main benefit of having conducted a new high school survey every 10 years is that the designers of each new survey are able to learn from mistakes made in the previous survey. The HSB improved on the NLS:72, and NELS improved considerably on the HSB. Prospects did not benefit much from the mistakes made in collecting the Sustaining Effects data because the Sustaining Effects data were never widely available enough to be subjected to close scrutiny. (The designers of Prospects did benefit, however, from improvements in the high school surveys—NELS, in particular.)

Lower Year-to-year Correlations

Two empirical facts illustrate why we need to focus on early schooling if we want to understand ethnic differences in achievement. First, the correlations between students' scores in one year and their scores in the following year tend to be lower during elementary school than during high school, even after correcting for measurement error (see table 4). In fact, year-to-year correlations are so high during high school (0.98 for both reading and math after correcting for measurement error) that hardly any students change their relative rank between the eighth and the twelfth grades. The lower year-to-year correlations during elementary school imply that cognitive skills are most malleable among young children and suggest that interventions aimed at changing students' skills may be most successful in the early school years.

Table 4. Year-to-year Correlations in Prospects and NELS

	Observed Correlation	True Correlation
Prospects		
Grade 1, Fall to Spring		
Reading	.49	.64
Math	.67	.81
Grade 3, Spring to Spring		
Reading	.69	.75
Math	.71	.80
NELS		
Grade 8, Spring to Spring		
Reading	.90	.98
Math	.94	.98

NOTE: Correlations are based on weighted data. They are disattenuated using the reliabilities reported in the Prospects Interim Report (1993) and the Psychometric Report for the NELS: 88 Base Year Through Second Follow-Up.

Lower Gain-Initial Score Correlations

A second, and related, reason to study achievement during elementary school is that the correlation between children's initial skills and how much they learn as they age is lower in elementary school than in high school. Everyday empiricism often suggests that additional advantages typically befall those who are already most advantaged. In terms of cognitive skills, this "fan spread" theory implies that students who have the best reading and math skills when they begin school will tend to gain the most reading and math skills as they move through school. Conversely, students who begin school with the fewest reading and math skills will tend to gain the fewest skills as they move through school. As students' skills diverge, their growth trajectories will come to resemble an opening fan.[13]

Scholars have debated for decades whether fan spread applies to test scores. For fan spread to occur, students' test score gains have to be positively correlated with their initial scores. More than 30 years ago, Benjamin Bloom (1964) argued in his famous study, *Stability and Change in Human Characteristics*, that gains on a wide variety of tests were uncorrelated with initial scores. But many scholars dismissed his findings as a result of measurement error in the tests (see, for example, Werts and Hilton 1977).[14] David Rogosa and John Willett (1985) have shown, however, that the relationship between gains and initial scores depends both on the shape of individual growth curves for a particular skill and on the age at which researchers measure initial status on that skill. The "true" correlation between initial status and gains can therefore be negative, positive, or 0, depending on what the particular test measures and when the children taking the test normally learn the skills that the test measures.

Table 5 shows that, even after correcting for measurement error, the correlations between gains and initial scores tend to be lower among elementary school students than high school students. The true gain-initial score correlations of around 0.50 during the first two years of high school indicate that the

[13] The fan-spread phenomenon is also known as the Matthew effect, after the biblical "For he that hath, to him shall be given… but he that hath not, from him shall be taken away…."

[14] Measurement error often creates a negative correlation between gains and initial scores. This is because, if we measure gains by subtracting observed Time 1 scores from observed Time 2 scores, the random error term in the Time 1 score appears with a positive sign in the Time 1 score, but with a negative sign in the gain score.

Table 5. Correlations between Initial Scores and Gains in Prospects and NELS

	Observed Gain, Initial Score Correlation	True Gain, Initial Score Correlation
Prospects		
Fall of Grade 1 to Spring of Grade 2		
Reading	-.33	-.14
Math	-.43	-.30
Spring of Grade 3 to Spring of Grade 5		
Reading	-.49	-.45
Math	-.43	-.36
NELS		
Spring of Grade 8 to Spring of Grade 10		
Reading	-.09	.53
Math	.06	.48
Spring of Grade 10 to Spring of Grade 12		
Reading	-.28	-.12
Math	-.08	.23

NOTE: Correlations are based on weighted data. They are disattenuated using the reliabilities reported in the Prospects Interim Report (1993) and the Psychometric Report for the NELS:88 Base Year Through Second Follow-Up and equations 11 and 13 in Willett (1988).

highest-scoring students at the end of eighth grade tend to gain the most during ninth and tenth grade. This is, of course, exactly what we would expect if students were tracked based on their math and reading scores and if students in higher-level math and English courses learned more math and reading skills. But these results contrast markedly with those for elementary school.[15] The correlations between true gains and true initial scores in Prospects tend to be negative. This implies that students who start first grade with higher scores tend to learn less between first and third grade than those who start first grade with lower scores.

[15] Note, however, that correlations between gains and initial scores are lower during the last two years of high school, especially for reading. Jencks and Phillips (1999) found near zero correlations between gains and initial scores in the High School and Beyond (HS&B) during the last two years of high school. If the NELS and HS&B tests measured all the skills that high-scoring students learn in the last two years of high school (e.g., calculus), the gain-initial score correlations might be as large and positive as they are for the first two years of NELS.

The negative correlation between gains and initial scores in Prospects suggests that elementary schools may be structured (intentionally or not) to help those children who have the weakest academic skills.[16] For example, first grade teachers may focus more on teaching children how to read than on improving the reading skills of those who already know how to read. Or perhaps elementary school math textbooks are geared to below average students at each grade level, so that students with the weakest math skills are challenged the most. Of course, reforms aimed at enriching the elementary school curriculum may end up improving learning among the most highly skilled students, thereby increasing the gap between high and low scorers.

Prospects is not the only data set to suggest that elementary schools may reduce rather than exacerbate the test score gap. Entwisle and Alexander have reported similar results for the BSS (for a recent example, see Alexander and Entwisle 1998). Because the black-white test score gap seems to widen more during elementary school than during high school, because children's scores are less fixed during elementary school than during high school, and because elementary schools may already help reduce the gap between high and low scorers, studies involving elementary school-age children are essential for understanding the development of ethnic differences in achievement. Fortunately, NCES is funding the Early Childhood Longitudinal Study (ECLS-K), which began collecting data on a national sample of kindergartners in the fall of 1998. The ECLS-K is much needed. One aspect of its design, however, will severely limit its contribution to our knowledge about ethnic differences in academic achievement.

Measuring Summer Learning

The ECLS-K currently plans to test a 25 percent subsample of students in the fall of first grade. For these children, the ECLS-K will be able to distinguish learning during kindergarten and first grade from learning during the summer

[16] An alternative explanation for the low gain-initial score correlation in elementary school is that the elementary school tests do not measure the skills that high-scoring students acquire during elementary school. This is almost certainly true. Yet, it must be more true for tests administered during elementary school than for tests administered during high school in order for it to explain why the gain-initial score correlations are so much lower in elementary school than in high school. If elementary school children "outgrew" the tests faster than high school students, then the lower gain-initial score correlations would be a by-product of not measuring high-scoring students' skills, but such a result would manifest itself as a ceiling effect on the elementary school tests. The Prospects tests used in the analysis in table 5 do not, however, show larger ceiling effects than the NELS tests.

following kindergarten. Yet learning during all other grades will not be separable into school year and summer components. This is a mistake because a number of studies, including the national Sustaining Effects study of the 1970s, suggest that the black-white gap widens over the course of schooling primarily because African American children gain less than white children during the summer (Entwisle and Alexander 1992, 1994; Hemenway et al. 1978; Heyns 1978, 1987; Klibanoff and Haggart 1981).[17] The only recent national data that speak to this question come from a nonrandom subset of Prospects students.[18] First, consider what we would conclude if we had measured black and white first graders' learning only at the beginning of first grade and the beginning of second grade. The first two columns of table 6 show that black and white children gain about the same skills in reading, vocabulary, and math over the course of a year. But dividing this yearly interval into learning that occurs during the winter (when school is in session) and learning that occurs during the summer (when school is not in session) yields a strikingly different picture of black-white differences in learning. The results show that, when school is in session, black first graders gain more reading and vocabulary skills than white first graders. Among students who start first grade with the same skills, blacks gain more vocabulary skills than, and about the same reading skills as, whites. But during the summer following first grade, blacks gain fewer reading and vocabulary skills than whites, both overall and among children who had similar skills at the end of first grade. The results for math show a similar but weaker pattern. These results resemble those from other studies (e.g., Heyns 1978, Entwisle and Alexander 1992, 1994).[19] Suppose we believe Phillips, Crouse, and Ralph's (1998) estimate that about half of the black-white gap is attributable to what occurs between first and twelfth grade. Without data on summer learning, we will never know how much of the black-

[17] The Sustaining Effects study was the first national longitudinal study to assess children's reading and math growth. In 1976, Sustaining Effects collected data on 120,000 students in over 300 public schools throughout the country (Carter 1983). It then followed a subsample of these students for three years, administering CTBS reading and math tests in the fall and spring of each school year as long as the students remained in the same schools (Won, Bear, and Hoepfner 1982).

[18] See appendix table A (page 132) for a comparison between this subsample and the larger cross-sectional sample. The subsample is quite advantaged relative to the original sample for reasons that remain a mystery to me.

[19] When Cooper and colleagues (1996) conducted a meta-analysis of the effect of summer vacation on test scores, they did not find consistent racial differences in summer learning. They did, however, find that middle-class children gained more in reading over the summer than lower-class children did. Cooper and colleagues explained that the lack of an effect of race was probably attributable to the fact that most of the studies in their meta-analytic sample partially controlled for family income before examining race differences.

Table 6. Black-White Differences in Reading, Vocabulary, and Math Growth on the CTBS in Prospects: School Year and Summer Comparisons

	Yearly Growth (fall, grade 1 to fall, grade 2)		Winter Growth (fall, grade 1 to spring, grade 1)		Summer Growth (spring, grade 1 to fall, grade 2)	
	Raw	Residualized	Raw	Residualized	Raw	Residualized
Reading Comprehension						
African American gain relative to European American	4.86	-4.97	19.59*	9.27	-14.73*	-20.47***
gain	(7.18)	(4.73)	(8.47)	(6.03)	(6.73)	(5.09)
Vocabulary						
African American gain relative to European American	13.67	5.46	24.38*	15.63*	-10.71**	-9.77*
gain	(10.22)	(5.21)	(10.68)	(6.37)	(3.66)	(3.32)
Math Concepts and Applications						
African American gain relative to European American	5.51	-3.13	11.26	8.62	-5.75	-18.75***
gain	(8.33)	(6.21)	(9.37)	(6.77)	(3.88)	(5.11)

NOTE: N=1,097. Numbers are unstandardized coefficients. Standard errors are in parentheses. Residual gain equations are errors-in-variables regressions that also control prior reading comprehension, vocabulary, and math concepts scores, corrected for measurement error using the reliabilities published in the Prospects Interim Report (1993). All equations are weighted using the 1993 weight, and all the standard errors are corrected for nonindependence within school districts. All equations also include gender and dummies for Asian American, Mexican American, Puerto Rican American, and other Hispanic students. See Phillips (1998) for more details on the sample and for results for other ethnic groups.

white gap is attributable to what occurs during the school year and how much is attributable to what occurs during the summer.

We typically design new surveys based on sparse and contradictory evidence, but the data on summer learning are too consistent to be ignored.[20]

[20] Nonetheless, the summer learning results still leave a number of important questions unanswered. We still do not know exactly how much of the widening of the black-white reading gap occurs during the summer as opposed to the school year. Nor do we know whether summers in early elementary school or later elementary school are most detrimental to black children's learning. Nor do we know how the relative contributions of families, schools, and neighborhoods to students' achievement change between the school year and summer vacation.

Summer vacations create a natural experiment that helps us begin to separate the effects of schooling from the effects of families and neighborhoods. Taking advantage of this experiment requires that we assess students twice a year: once in the fall and once in the spring. Testing students twice a year is expensive and imposes a considerable burden on students and schools, but testing students only once a year or once every two years is largely a waste of resources because it confounds what are potentially separable causes of differences in children's learning. A better alternative would be to randomly select multiple subsamples of students who would be tested twice a year. In the case of the ECLS-K, for example, one random subsample could be tested in the fall of first grade, another in the fall of second grade, another in the fall of third grade, and so on. This would minimize the burden to any particular student or school while maximizing our knowledge about how much of the learning gap arises during the summer.

Maximizing Measurement Variation within Data Sets

Another way to improve surveys of young children's academic growth is to increase the measurement variation within each survey. We can do this in the following two ways: by including multiple measures of the same skill and by including precise measures of each particular set of skills.

Multiple Measures of the Same Skill

Other than the Prospects survey, the only other contemporary national survey that has tested young children multiple times is the Children of the National Longitudinal Survey of Youth (CNLSY). Unlike school surveys, which have a grade-based sampling design, the CNLSY is age-based. The Bureau of Labor Statistics (BLS) first funded a survey of the mothers of these children in 1979, when the prospective mothers were 14- to 22-years-old. It has resurveyed them every year since then. In 1986, the BLS began collecting data on all the children who had been born to the original sample of mothers. These children, as well as all additional children born to the mothers, have been followed at two-year intervals since then.

The CNLSY administers math, reading, and vocabulary tests to all children who are at least five years old. These data show a small black-white *reading* gap (0.20 standard deviations) among 5-year-olds, but they show a large black-white *vocabulary* gap (0.98 standard deviations) among the very same children. If the CNLSY had administered more than one type of reading test and more than one type of vocabulary test to these children, we would be able to determine whether the stark difference in the size of these gaps reflects idiosyncrasies of the tests or a real phenomenon.

Precise Measures of Each Type of Skill

A survey can also enhance our understanding of ethnic differences in test scores by measuring a particular set of skills as precisely as possible. For example, the CTBS total math test administered by Prospects is composed of math computation and math concepts subtests. Analyzing these subtests separately reveals potentially important differences that we would miss if we only analyzed scores on the total math test. Table 7 shows ethnic differences in third graders' scores on the concepts and computation tests, as well as ethnic differences in gains on these tests between the third and sixth grades. The first panel shows that the gap between African American and white third graders is almost twice as large on the math concepts test (0.95 standard deviations) as on the math computation test (0.51 standard deviations). The same is true for the gap between whites and Mexican Americans. In contrast, Asian American third graders score at about the same level as white third graders on the math concepts test, but they score 0.65 standard deviations higher than whites on the math computation test.

The second panel in table 7 shows that African American and Mexican American children's *gains* also differ across the subtests. Although African American and Mexican American students gain about the same amount as whites on the math computation test, they gain more than whites on the math concepts test. These results illustrate why surveys need to administer tests that measure a wide range of skills and why analysts need to examine subtests separately, even when they sound relatively similar in name or content.

Maximizing Variation between Surveys

No matter how hard one works to perfect every survey, data sets inevitably end up with idiosyncrasies that can affect the analytic results. The best way to ensure that these biases do not affect our policy decisions is to collect as much data as possible, using different samples, survey designs, contractors, tests, technical review panels, and so on. Then, when we combine the results from these various studies, robust relationships should persist across the

Table 7. Comparison of Math Concepts Growth and Math Computation Growth: Evidence from Prospects Third Grade Cohort

	Math Concepts and Applications		Math Computation	
	1	**2**	**1**	**2**
Score in spring of third grade				
Asian American	4.65	8.32+	23.21***	24.06***
	(4.42)	(4.31)	(3.76)	(3.71)
	.10	*.17*	*.65*	*.68*
African American	-45.35***	-39.01***	-18.22***	-17.50***
	(2.66)	(2.60)	(2.28)	(2.25)
	-.95	*-.82*	*-.51*	*-.49*
Mexican American	-37.68***	-21.93***	-13.25***	-4.86+
	(3.07)	(3.10)	(2.67)	(2.71)
	-.79	*-.46*	*-.37*	*-.14*
Intercept	694.53***	679.25***	677.04***	672.32***
	(2.03)	(3.84)	(1.72)	(3.31)
Yearly growth between third and sixth grade				
Asian American	6.31***	4.65***	3.85**	3.73*
	(1.39)	(1.39)	(1.47)	(1.47)
	.36	*.27*	*.16*	*.15*
African American	2.75**	2.34**	.30	1.50
	(.87)	(.88)	(.91)	(.92)
	.16	*.14*	*.01*	*.06*
Score in spring of third grade				
Mexican American	4.66***	3.74**	-.48	.65
	(1.05)	(1.07)	(1.10)	(1.12)
	.27	.22	.02	.03
Grade	17.44***	19.34***	26.97***	25.06***
	(.80)	(1.40)	(.82)	(1.46)
Pseudo-R^2 for Intercept	.16	.25	.06	.15
Pseudo-R^2 for Slope	.04	.07	.03	.06

NOTE: N=4,550. Numbers are unstandardized coefficients. Standard errors are in parentheses. Italicized numbers are standardized coefficients for the intercept equation and proportions of the average gain for the slope equation. Equation 1 also includes gender and dummies for other ethnic groups on both the intercept and slope, and a nonlinear grade/age term on the slope. In addition, equation 2 includes dummies for mother's education, dummies for region, and dummies for urbanism on both the intercept and slope. All estimates come from weighted 2-level hierarchical models, with grade at level 1 and students at level 2. Standard errors and significance tests are corrected for clustering within districts by inflating the standard errors by the ratio of the standard errors produced by an unweighted 3-level model and an unweighted 2-level model. See Phillips (1998) for more details.

studies, despite their differences.[21] The best way to illustrate the importance of analyzing multiple surveys is to compare data from surveys that should, in theory, yield similar results.

For both the CNLSY and Prospects, I estimated growth models in which I predicted ethnic differences in students' initial test scores and learning rates.[22] The first panel of table 8 compares ethnic differences in initial reading scores in Prospects and the CNLSY. The CNLSY suggests that African American and white 5-year-olds have nearly identical reading recognition scores. (This finding resembles Entwisle and Alexander's BSS reading results for first graders in Baltimore.) Yet the results from Prospects show a 0.87 standard deviation gap in reading comprehension among first graders. Holding mothers' education, region of residence, and urbanism constant, this gap shrinks to 0.78 SDs.[23] (These results resemble those from the Sustaining Effects study.) If the CNLSY or Prospects had administered more than one reading test to their respondents, we would be able to determine whether the small black-white reading gap in the CNLSY is attributable to sample differences between the CNLSY and Prospects or to differences between the PIAT and CTBS tests.

[21] The same principle underlies meta-analysis. See Cook (1993) for a discussion of the principle of "heterogeneous irrelevancies." Of course, combining studies will not wash out biases that go in the same direction. For example, student mobility creates an upward bias in longitudinal studies because students who change schools or districts (and are therefore not retested) tend to have lower scores than students who stay in the same schools. This creates a major problem for researchers and policymakers who are often most interested in what happens to the most disadvantaged students.

[22] Children in the CNLSY first took Peabody Individual Achievement Tests (PIAT) in reading and math when they were 5 years old. I measured their learning rates in age-in-months. Children in Prospects first took Comprehensive Tests of Basic Skills (CTBS) in reading in the fall of first grade. I measured their growth rates in years. For details on the samples and analysis, see Phillips (1998).

[23] The contradiction between the reading results in Prospects and those in the CNLSY does not seem to be attributable to the fact that the CNLSY administered a reading recognition test, while Prospects administered a reading comprehension test. Equations using the CNLSY reading comprehension test as the dependent variable (not shown) produced results nearly identical to those for reading recognition. See Phillips (1998) for a discussion of whether the contradictory results are attributable to the psychometric properties of the different tests. She concludes that the kurtosis of the CNLSY tests probably does not account for the different results.

Table 8. Ethnic Differences in Reading Growth: Comparison of the CNLSY and Prospect

	CNLSY		Prospects	
	1	**2**	**1**	**2**
Reading score at age 5 (in fall, grade 1)				
African American	-.031 (.097)	.030 (.100)	-.868*** (.058)	-.776*** (.057)
Mexican American	-.552*** (.156)	-.430** (.164)	-.650*** (.087)	-.382*** (.088)
Linear reading growth in years				
African American	-.240*** (.024) *.121*	-.240*** (.024) *.120*	-.107** (.026) *.049*	-.071** (.025) *.034*
Mexican American	-.132*** (.041) *.070*	-.084+ (.048) *.042*	-.020 (.038) *.009*	.048 (.039) *.023*
Pseudo-R^2 for Intercept	.02	.12	.12	.20
Pseudo-R^2 for Slope	.10	.14	.04	.12

NOTE: To facilitate the comparison across data sets, all coefficients and standard errors are expressed as a proportion of the overall standard deviation of students' initial scores. In addition, italicized numbers in the gain equation express the gain as a proportion of the average gain in a particular data set. Moreover, the monthly gain coefficients and standard errors in the CNLSY have been multiplied by 12 to approximate the yearly gain interval in Prospects. Equation 1 includes gender and dummies for other ethnic groups on both the intercept and slope, and a nonlinear grade/age term on the slope. In addition, equation 2 includes dummies for mother's education, dummies for region, and dummies for urbanism on both the intercept and slope. All estimates come from weighted 2-level hierarchical models, with age/grade at level 1 and students at level 2. Standard errors and significance tests are corrected for the nonindependence of siblings (in the CNLSY) and for clustering within districts (in Prospects) by inflating the standard errors by the ratio of the standard errors produced by an unweighted 3-level model and an unweighted 2-level model. Prospects N=4,647; CNLSY N=2,153. See Phillips (1998) for more details.

Surprisingly, the CNLSY and Prospects results for Mexican Americans' initial reading skills are much more consistent.[24] The CNLSY suggests that Mexican American and white 5-year-olds' reading recognition skills differ by about 0.55 standard deviations. Prospects suggests that Mexican American and white first graders' reading comprehension skills differ by about 0.65 standard

[24] The similarity of these estimates is surprising in light of the fact that the CNLSY does not include children of recent immigrants. One would expect this sampling difference to affect comparisons of Mexican Americans' scores more than it affects comparisons of African Americans' scores, but that does not seem to be the case.

deviations. Among Prospects first graders whose mothers have the same amount of schooling and who live in the same region, the Mexican American-white reading comprehension gap shrinks by almost half—to 0.38 SDs. The reduction is not nearly so large in the CNLSY, but the gap after controlling mothers' education and region is remarkably similar (0.43 SDs).

The bottom panel of table 8 compares test score *growth* in the CNLSY and Prospects. Both the CNLSY and Prospects suggest that African American children learn fewer reading skills than white children during the early elementary school years, but the learning gap is less extreme in Prospects than in the CNLSY. In the CNLSY, African American children gain about 1 raw score point (over a fifth of the age-5 standard deviation) less than white children, each year. This is about 12 percent less than the average gain. Between the first and third grades, African American children in Prospects gain only 5.6 fewer points per year than whites, which is 5 percent less than the average gain and 11 percent of the first grade standard deviation.

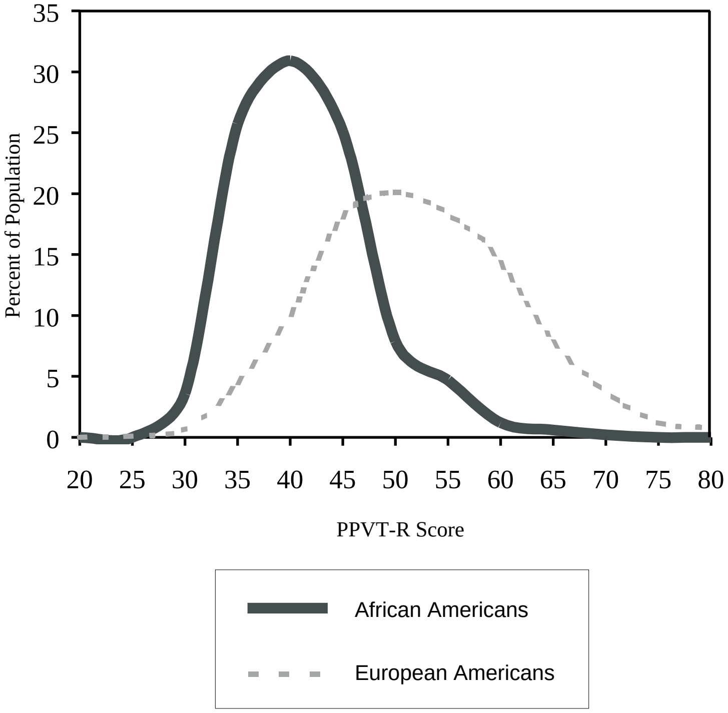
The CNLSY and Prospects also tell somewhat different stories about the relative reading trajectories of Mexican Americans and whites. In the CNLSY, Mexican American children gain about two-thirds of a raw score point (about 0.13 age-5 standard deviations) less than white children each year, which is equivalent to a gain of 7 percent less than average. This difference shrinks to a gain of 4 percent less than the average among white and Mexican American children whose mothers have the same amount of schooling and who live in the same region of the country. In contrast, Prospects suggests that young Mexican American children gain about the same reading skills as young white children, especially if we compare children whose mothers have the same amount of schooling and who live in the same region of the country.

The results in table 8 illustrate the difficulty of replicating results across surveys. They also serve as a cautionary tale to researchers who analyze data from a single source. Finally, these results underscore the problem of having only two contemporary national longitudinal data sets with which to describe elementary school children's academic development.

Studying Education Prior to Formal Schooling

Another reason we know so little about when ethnic differences arise and how they change with age is that NCES has, until recently, mostly ignored education that occurs outside of formal institutions. Although teachers have long argued that families exert the most influence on children's academic skills, few educational researchers study academic achievement before children enter elementary school. This habit hampers our understanding of ethnic differences in achievement because at least half of the black-white gap that exists at the end of twelfth grade can be traced to the gap that already existed at the beginning of first grade. Data from the CNLSY show that we can trace the vocabulary gap back to when black and white children are three years old (see figure 3). If

Figure 3. Vocabulary Scores for Black and White 3-Year-Olds, 1986–94



SOURCE: National Longitudinal Survey of Youth Child Data, 1986-94. Black N=507; white N=949. Figure is based on black and white 3-year-olds who took the Peabody Picture Vocabulary Test-Revised (PPVT-R). The scores shown are the standardized residuals, coded to a mean of 50 and a standard deviation of 10, from a weighted regression of children's raw scores on their age in months, age in months squared, and year-of-testing dummies. Lines are smoothed.

the CNLSY had measured infants' and toddlers' cognitive skills, we might be able to trace the gap back even farther.[25]

Focusing on family influences, as opposed to school influences, is not without its disadvantages, of course. Although the CNLSY does a better job than education surveys of measuring children's experiences outside of school, it does a considerably worse job of measuring their experiences inside school. Combining the advantages of a survey like Prospects with those of a survey like the CNLSY must become commonplace during the next century if we are to make any progress on the test score gap issue.

Fortunately, NCES is planning such a survey, in a joint effort with the National Center for Health Statistics (NCHS), the National Institutes for Child Health and Human Development (NICHD), the Administration for Children, Youth, and Families (ACYF), and the U.S. Department of Agriculture. The ECLS-Birth cohort study, targeted to begin in the year 2000, plans to follow a cohort of 6-month-olds through the end of first grade.

Hopefully, the designers of this study will learn from the strengths and limitations of the CNLSY. Important strengths of the CNLSY include sampling siblings in order to estimate family background effects, testing mothers' cognitive skills (testing mothers and fathers would be even better), and measuring parenting practices using both self-reports and interviewers' observations (observing parenting at more frequent intervals and using time diaries to collect parenting data would be even better). Limitations of the CNLSY include not trying to measure children's cognitive skills prior to age 3 and administering only one type of reading and vocabulary test to children.

Surveys versus Experiments

Studying elementary school students, assessing at least some of these students in the fall and spring of every school year, using multiple tests to measure students' reading and math skills, funding multiple surveys of chil-

[25] Because children's cognitive skills are moderately stable between infancy and early childhood, ethnic differences may be as well. Measures of infant habituation and recognition memory correlate 0.36, on average, with childhood IQ (see the meta-analysis by McCall and Carriger 1993). Thompson, Fagan, and Fulker (1991) also find that visual novelty preference at 5 to 7 months of age is associated with both IQ and achievement at age 3, and Dougherty and Haith (1997) find that visual anticipation and visual reaction time at 3.5 months are associated with IQ at age 4.

dren, and studying educational development before formal schooling begins—all would help improve our understanding of how ethnic differences in academic achievement change with age.[26] These design changes would also help us generate better theories about the causes of ethnic differences in academic skills.[27] Regardless of how we decide to design our national surveys, however, the best survey data will not tell us how to raise children's achievement—neither will

[26] An important methodological issue that I have not discussed here is the problem of choosing the correct metric with which to measure academic growth. Because the metric issue is so perplexing, almost all researchers simply use the particular test at their disposal, without questioning how the test's metric affects the results. For instance, when black and white children gain 50 points on a vocabulary test scored using IRT, we typically assume that black and white children *learned* the same amount of vocabulary over that interval. But even IRT-scored tests may not have interval-scale properties, which means that a gain from 250 to 300 may not be equivalent to a gain from 350 to 400. The only solution I see to the problem of determining whether gains from different points on a scale are equivalent is to associate a particular test with an outcome we want to predict (say, educational attainment or earnings), estimate the functional form of this relationship, and then use this functional form to assess the magnitude of gains. For example, if test scores are linearly related to years of schooling, then gains of 50 points can be considered equal, regardless of the starting point. If the *log* of scores is linearly related to years of schooling, however, then a gain of 50 points from a lower initial score is worth more than a gain of 50 points from a higher initial score. This "solution" is, of course, very unsatisfactory, because the functional form of the relationship between test scores and outcomes undoubtedly varies across outcomes.

[27] We should, however, question the assumption that large national studies should be preferred over multiple local studies. Ethnic differences in test scores vary across states, school districts, and schools. This means that national surveys probably mask much of the differential development in academic skills that occurs in particular school districts throughout the country. Scholars and policymakers should begin to debate the financial, political, and quality trade-offs between formally selecting nationally representative samples and purposefully selecting a large number of locally representative samples that are informally representative of the types of students to whom we would like to generalize nationally (see Cook [1993] on the logic of purposive sampling based on the "principle of proximal similarity"). Over the past few decades, we have learned the most about the correlates of young students' academic development from Entwisle and Alexander's BSS study. This may be because NCES never put the same effort into a national study of young children's achievement that it put into NELS. Or it may be that local surveys, run by university researchers who are better able to generate goodwill among local school administrators, teachers, and parents, are more effective than national surveys. In either case, a potential alternative to funding several large national surveys might be for NCES, in collaboration with private foundations and state and local governments, to support a large number of local longitudinal studies, on the condition that the procedures and measures be comparable enough to combine results across sites.

the best longitudinal methods.[28] Our ultimate goal in collecting data on student achievement is presumably to raise all children's achievement while reducing variation in achievement, not just to produce long descriptive reports or fill academic journals. If that is our goal, the standards that we need for assessing causality are much higher than survey data can satisfy.

The best way to learn what will reduce ethnic differences in achievement is to conduct randomized field trials.[29] Such studies could include programs designed by researchers, based on theories generated from nonexperimental data, as well as programs designed by teachers and administrators, based on programs that already seem to have worked on a small scale. The recent turn to natural experiments in economics and sociology may also help us begin to identify the causs of ethnic differences in academic achievement.[30]

We must also remember that it is not logically necessary to understand the causes of a social problem before successfully intervening to fix it. Instead of starting with the question "What causes the gap?" and hoping that the answers will lead to effective interventions, we may need to instead start with the question "How can we reduce the gap?" and then collect the kind of information (namely, experimental data) that will enable us to do just that.

[28] Methods for measuring longitudinal change and easy-to-use software packages for implementing these methods have proliferated over the past decade. Although these methods are often very helpful for describing longitudinal change and its correlates, they do not increase our ability to make correct causal inferences—nor are these methods always the best analytic choice. These new methods need to be subjected to cross-disciplinary discussions about their costs and benefits. The typical user neither understands the main advantages of multilevel models nor knows that other procedures (such as fixed effects models with a correction for the loss of degrees of freedom resulting from the intraclass correlation) have some of the same advantages without some of the disadvantages. See McCallum et al. (1997) for a review of multilevel methods for describing growth (including structural equation models) and Kreft (1996) for a discussion of the advantages and disadvantages of using random coefficient models.

[29] See Nave, Miech, and Mosteller (1998) for a review of the use of random field studies in education and for an argument in favor of funding more of them.

[30] When estimating the "effect" of educational processes on students' achievement gains, more frequent attention to correcting for measurement error in initial test scores would also help. In nonexperimental studies that find a positive effect of school or family characteristics on students' learning, the most frequent threat to validity is that these school or family characteristics mainly serve as proxies for initial skills that were imperfectly measured.

References

Alexander, K. L., and Entwisle, D. R. (1998). *Isolating the School's Contribution to Achievement: School Year Versus Summer Gains.* Paper presented at the annual meeting of the American Association for the Advancement of Science.

Ainsworth-Darnell, J. W., and Downey, D. B. (1998). Assessing the Oppositional Culture Explanation for Racial/Ethnic Differences in School Performance. *American Sociological Review 63*: 536–53.

Bloom, B. (1964). *Stability and Change in Human Characteristics*. New York: Wiley.

Carter, L. F. (1983). *A Study of Compensatory and Elementary Education: The Sustaining Effects Study. Final Report.* Santa Monica, CA: System Development Corporation.

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, P. J., and Ludwig, J. (1997). Weighing the Burden of "Acting White": Are There Racial Differences in Attitudes toward Education? *Journal of Policy Analysis and Management 16*(2): 656–78.

Cook, T. D. (1993). A Quasi-Sampling Theory of the Generalization of Causal Relationships. *New Directions for Program Evaluation 57*: 39–81.

Cooper, H., Nye, B., Charlton, K., Lindsay, J. and Greathouse, S. (1996, fall). The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-analytic Review. *Review of Educational Research 66*: 227–268.

Cooper, H., and Hedges, L. (Eds.). (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Dougherty, T. M., and Haith, M. M. (1997, January). Infant Expectations and Reaction Time as Predictors of Childhood Speed of Processing and IQ. *Developmental Psychology 33*: 146–155.

Entwisle, D. R., and Alexander, K. L. (1988, May). Factors Affecting Achievement Test Scores and Marks of Black and White First Graders. *The Elementary School Journal 88*: 449–471.

Entwisle, D. R., and Alexander, K. L. (1990). Beginning School Math Competence: Minority and Majority Comparisons. *Child Development 61*: 454–71.

Entwisle, D. R., and Alexander, K. L. (1992, February). Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School. *American Sociological Review 57*: 72–84.

Entwisle, D. R., and Alexander, K. L. (1994, June). Winter Setback: The Racial Composition of Schools and Learning to Read. *American Sociological Review 59*: 446–460.

Fordham, S., and Ogbu, J. (1986). Black Students' School Success: Coping with the Burden of "Acting White." *Urban Review 18*(3): 176–206.

Hemenway, J. A., Wang, M., Kenoyer, C. E., Hoepfner, R., Bear, M. B., and Smith, G. (1978). *Report #9: The Measures and Variables in the Sustaining Effects Study.* Santa Monica, CA: System Development Corporation.

Heyns, B. (1978). *Summer Learning and the Effects of Schooling*. New York: Academic Press.

Heyns, B. (1987). Schooling and Cognitive Development: Is There a Season for Learning? *Child Development 58*: 1151–1160.

Jencks, C., and Phillips, M. (1999). Aptitude or Achievement: Why Do Test Scores Predict Educational Attainment and Earnings? In S. E. Mayer and P. E. Peterson (Eds.), E*arning and Learning: How Schools Matter (*pp. 15–47). Washington, DC: Brookings Institution Press and Russell Sage Foundation.

Jencks, C., and Phillips, M. (1998). The Black-White Test Score Gap: An Introduction. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 1–51). Washington, DC: Brookings Institution Press and Russell Sage Foundation.

Kao, G., Tienda, M., and Schneider, B. (1996). Racial and Ethnic Variation in Academic Performance. *Research in Sociology of Education and Socialization 11*: 263–297.

Klibanoff, L. S., and Haggart, S. A. (1981). *Report #8: Summer Growth and the Effectiveness of Summer School*. Mountain View, CA: RMC Research Corporation.

Kreft, I. C. G. (1996). *Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies.* Los Angeles, CA: California State University.

MacCallum, R. C., Kim, C., Malarkey, W. B., and Kiecold-Glaser, J. K. (1997). Studying Multivariate Change Using Multilevel Models and Latent Curve Models. *Multivariate Behavioral Research 32*(3): 215–253.

McCall, R. B., and Carriger, M. S. (1993, February). A Meta-analysis of Infant Habituation and Recognition Memory Performance as Predictors of Later IQ. *Child Development 64* (February): 57–79.

Nave, M., Meich, E. U., and Mosteller, F. (1998). *A Rare Design: The Role of Field Trials in Evaluating School Practices.* Paper presented at the American Academy of Arts and Sciences. Harvard University and the American Academy of Arts and Sciences. Cambridge, MA.

Ogbu, J. (1978). *Minority Education and Caste: The American System in Cross-cultural Perspective.* New York: Academic Press.

Phillips, M. (1998). *Early Inequalities: The Development of Ethnic Differences in Academic Achievement During Childhood*. Ph.D. dissertation, Northwestern University.

Phillips, M., Brooks-Gunn, J., Duncan, G., Klebanov, P., and Crane, J. (1998). Family Background, Parenting Practices, and Test Performance. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 103–145). Washington, DC: Brookings Institution Press.

Phillips, M., Crouse, J., and Ralph, J. (1998). Does the Black-White Test Score Gap Widen After Children Enter School? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 229–272). Washington, DC: Brookings Institution Press.

Pindyck, R. S., and Rubinfeld, D. L. (1991). *Econometric Models and Economic Forecasts* (3rd ed.). New York: McGraw-Hill.

Rogosa, D. A., and Willett, J. B. (1985). Understanding Correlates of Change by Modeling Individual Differences in Growth. *Psychometrika 50* (2): 203–228.

Thompson, L. A., Fagan, J. F., and Fulker, D. W. (1991). Longitudinal Prediction of Specific Cognitive Abilities from Infant Novelty Preference. *Child Development 62*: 530–538.

Werts, C. E., and Hilton, T. L. (1977). Intellectual Status and Intellectual Growth, Again. *American Educational Research Journal 14*(2): 137–146.

Willett, J. B. (1988). Questions and Answers in the Measurement of Change. In E. Z. Rothkopf (Ed.), *Review of Research in Education* (pp. 345–422). Washington, DC: American Educational Research Association.

Won, E. Y. T., Bear, M. B., and Hoepfner, R. (1982, January). *Background, Schooling, and Achievement.* Technical Report 20 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills. Santa Monica, CA: System Development Corporation.

Table A. Descriptive Statistics for Prospects Summer Sample and Comparison with Cross-sectional Sample

	First Grade Cohort						
	Longitudinal			Cross-sectional			SD
Variables	Mean	SD	N	Mean	SD	N	Diff.
Reading comprehension	486.62	62.90	1,097	478.52	67.81	9,422	.12
Vocabulary	494.78	63.51	1,097	481.71	62.87	9,408	.21
Math concepts and applications	500.18	64.35	1,097	481.08	68.73	9,237	.28
Male	.51	.50	1,097	.51	.50	11,349	.00
Asian American	.01	.12	1,097	.03	.17	11,357	-.12
European American	.79	.40	1,097	.68	.46	11,357	.24
African American	.11	.31	1,097	.15	.36	11,357	-.11
Mexican American	.07	.25	1,097	.09	.29	11,357	-.07
Puerto Rican Am.	.01	.09	1,097	.01	.09	11,357	.00
Other Latino	.01	.10	1,097	.03	.18	11,357	-.11
Mom's educ in yrs.	13.29	2.00	1,097	12.92	2.10	10,529	.18
Mom is hs. dropout	.10	.30	1,097	.16	.37	10,529	-.16
Mom is hs. grad.	.23	.42	1,097	.28	.45	10,529	-.11
Mom has some coll.	.49	.50	1,097	.40	.49	10,529	.18
Mom is coll. grad.	.18	.38	1,097	.16	.36	10,529	.06
Live in north	.11	.31	1,097	.19	.39	11,357	-.21
Live in midwest	.27	.44	1,097	.18	.39	11,357	.23
Live in south	.48	.50	1,097	.38	.49	11,357	.20
Live in west	.15	.36	1,097	.24	.43	11,357	-.21
Live in urban area	.07	.26	1,097	.25	.43	11,357	-.42
Live in rural area	.49	.50	1,097	.38	.48	11,357	.23
Live in suburban area	.44	.5	1,097	.38	.48	11,357	.13

NOTE: Longitudinal sample includes children with valid original scale scores in fall of grade 1, spring of grade 1, and fall of grade 2, ethnicity, gender, mother's education, region, and urbanism, and is weighted with the spring 1993 weight. Cross-sectional sample includes all children selected in the base-year sample and is weighted by the 1991 weight. I used the original scale scores for these analyses because the fall of grade 2 adjusted scale scores are incorrect.

Certification Test Scores, Teacher Quality, and Student Achievement[1]

Ronald F. Ferguson with Jordana Brown
Malcolm Wiener Center for Social Policy
John F. Kennedy School of Government
Harvard University

Raising student achievement levels in primary and secondary schools is again a top national priority. Campaigning politicians at every level of government are promising to improve education, but many of the measures they are proposing have not been firmly established by research to be effective. This paper concerns standardized testing of teachers. In the constellation of measures that might contribute to achievement gains, certification testing of teachers is one of many, and not necessarily the most important.[2] Nonetheless, certification testing, especially for new teacher applicants, is now used in 44 states. In 1980, only three states tested teacher candidates. By 1990, the number had catapulted to 42. Given that so many states are now involved in this broad-based national experiment, a serious effort to learn from it seems warranted. This paper reviews evidence on the relationship of teachers' test scores to student achievement and frames some of the questions that a serious, nationally coordinated program of research might address over the next decade.

[1] Work on this paper has benefited from the financial support of the Annie E. Casey Foundation, the John D. and Catherine T. MacArthur Foundation, and the Rockefeller Foundation, through their support for the National Community Development Policy Analysis Network Project on Education and Youth in Community Development. One section of the text comes from another paper by this author, "Can Schools Narrow the Black-White Test Score Gap?" published recently in Christopher Jencks and Meredith Phillips, (Eds.), *The Black-White Test Score Gap* (1998). That section appears here with the permission of Brookings Institution Press.

[2] For example, achieving and maintaining a high level of quality in teacher education and professional development programs should be key elements in any strategy to improve teacher quality.

Measuring Teacher Effectiveness

The challenge of measuring teacher effectiveness applies to both new and incumbent teachers; and, in at least a few cases, standardized teacher testing has been used for both.[3] However, unlike new teachers, experienced teachers have other ways of demonstrating their effectiveness. It seems reasonable to argue that incumbents should be judged on what they do in the classroom, not on a test score. A commonsense procedure is to have expert observers rate teachers on their classroom practice. Alternatively, districts can judge teachers based on measures of student achievement, such as test-score gains. Hence, we have three indicators of current and potential teacher effectiveness: (1) teachers' test scores, (2) observers' ratings of teachers' professional classroom practice, and (3) students' achievement gains.

However, each of these measures has notable flaws. Regarding the first, certification tests may not measure those aspects of teacher skill that matter most. Regarding the second, observers may not rate teachers on the practices that matter most, or they may make mistakes in recording what they observe. Even the third, students' test-score gains, is an imperfect measure of what we really want to know: *the teacher's contribution* to producing the gains. Because other factors such as student, home, school, and community characteristics affect achievement as well, teachers deserve neither all of the credit for successes nor all of the blame for failures.

Ideal assessments of teacher quality would involve directly measuring what teachers contribute to student learning. Unfortunately, since such measures are infeasible, we must resort to various approximations—typically, one or more of the three types listed above or estimates that use them in multivariate statistical analyses. In a standard multivariate analysis, the dependent variable is the student test score. Explanatory variables include school, family, student, and community characteristics, and often a baseline value of the student test score. When a binary (i.e., 0,1) indicator variable for each individual teacher

[3] Two states, Arkansas and Texas, have tested incumbent teachers as well as those just entering the profession. Policymakers in Massachusetts are currently considering whether to test its current teachers, and other states may follow suit.

is included in such an analysis, its estimated regression coefficient is a measure of the teacher's contribution to the test score.[4]

Using this method of estimation, new findings for both Texas and Tennessee indicate that the teacher a student has in a particular year affects learning gains a great deal, not only for the current year, but for the next several years as well.[5] Estimated productivity differences among teachers in these new studies are large[6] and remind us how very important it is to select and retain the most effective teacher candidates.[7] When the data are adequate and the analysis is done appropriately, the technique that produced these new findings is probably the best that we can do at measuring the effectiveness of individual teachers.

[4] Under certain conditions, such a coefficient may be biased upward or downward. The rank order among teachers may even be distorted, such that some teachers appear more effective than others when they are not. The degree to which such analyses produce mistakes depends on such things as the completeness of the data used to control for confounding factors and the appropriateness of the specific techniques used for estimation.

[5] During the 1990s, teams of researchers have assembled large new longitudinal data sets for Texas and Tennessee. They include more students and more information than any previous compilation. These data permit researchers to follow tens of thousands of individual children's progress, including achievement gains associated with individual teachers, across several grade levels. For Texas, see Rivkin, Hanushek, and Kain (1998) and Kain (1998). For Tennessee, see Sanders and Rivers (1996) and Sanders, Saxton, and Horn (1998). There is an ongoing debate about how important it is to include additional student background variables when estimating the effects of teachers on students' test score gains. We agree with Kain (1998) that student background variables are important to include and that the teacher effects estimated by Sanders and his coauthors in the studies cited here might change (though probably not dramatically) if such controls were included.

[6] Earlier studies of this type have also identified large differences in teacher effectiveness. See Hanushek (1986, 1992).

[7] Of course, the challenge could be stated more broadly, to include affecting the size and composition of the teacher applicant pool. That might include considering a broader range of policy alternatives such as pay scales and career prep strategies for students in grades K–12 to attract them toward teaching careers. It might also include consideration of screening and hiring practices, particularly those that induce districts to hire the most effective applicants. For discussions of both salary-related issues and hiring practices, see Ballou and Podgursky (1997) and Murnane et. al. (1991). Another challenge is to help less effective teachers to improve their skills and knowledge. See, for example, Darling-Hammond (1997) for a discussion of these issues from the National Commission on Teaching and America's Future.

Unfortunately, this type of analysis is usually impractical. Few states or districts are anywhere close to possessing the data needed to implement it well on a large scale. Even for states that have the data and the capacity, the method does not work for judging new teacher candidates and others for whom there are not several years of appropriate data. Hence, standardized competency tests and other observational methods of judging professional practice offer more feasible alternatives for making judgments or predictions about professional effectiveness.

Whether competency testing is generally superior to observational teacher-rating methods is an open question that warrants more attention. However, it seems clear that competency testing is less expensive than observational methods when there are thousands of teachers to be assessed. It also seems clear that observational assessments of teaching *candidates*, conducted under contrived conditions or during student teaching, are likely to be highly variable in their quality. Given these cost advantages and the lack of alternative methods that are reliable and consistent, standardized teacher competency tests may be the best way of measuring teacher quality we have when thousands (or tens of thousands) of teachers need to be assessed.

Do Teachers' Scores Predict Student Achievement?

Most teacher certification exams are vulnerable to a variety of fair criticisms. For example, many have not been well validated and admittedly measure only a small fraction of the skills that make teachers effective. Yet evidence from studies that actually estimate relationships between students' and teachers' scores, including my own work reviewed below, suggests generally that teachers' test scores *do* help in predicting their students' achievement.

When I first encountered this topic in the late 1980s, it came as a great surprise to me that predictive validity in the relationship of teachers' to students' scores was not among the criteria for validating teacher certification exams.[8] States all over the nation were using such tests to screen candidates in and out of the profession, with no firm evidence that scores predicted teaching effectiveness! This was astonishing, especially since passing rates for non-white candidates were substantially lower than for whites. If teachers' scores

[8] This is still the case.

were poor predictors of student performance, then there was a disproportionate impact on nonwhite candidates that was probably illegal.[9] On the other hand, if it turned out that teachers' scores were important predictors, then there was a trade-off between the interests of low-scoring teacher candidates versus children's right to a quality education. As Bernard R. Gifford, an African American and Dean of the Graduate School of Education at the University of California wrote in 1985:

> If we do have a commitment to quality education for all, as part of our dedication to the principles of equality, then we will not change the requirements to fit the present median performance of minority applicants to teacher education programs. Rather, we will keep the desired performance level and provide the kinds of support and training that will make it possible for minority applicants to garner the learning and experience needed to pass the examinations . . .
>
> To put forth the argument that minority youngsters, the most disadvantaged of the poor, and the least able to emancipate themselves from their impoverished surroundings, should be taught by our less-than-best teachers is to pervert the nature of justice. As admirable and important as is the goal of increasing the ranks of minority teachers, this objective must not be put before the more fundamental objective of securing good teaching for those who need it the most.[10]

In this passage, Gifford assumes that the exams measure skills that matter for predicting teacher effectiveness. Many others have disagreed with this assumption.[11] The lack of a well-organized body of evidence on how teacher

[9] Pressman and Gartner (1986, 11) went so far as to assert, "There is no evidence that what is being tested relates to the selection of persons who will be effective teachers." Their article also discusses some of the legal challenges that tried to stop the use of certification exams.

[10] See Gifford (1985, 61).

[11] For example, see Pressman and Gartner (1986) for a very skeptical discussion about competency testing.

characteristics (including race and test scores) relate to teacher effectiveness has fostered confusion in both scholarly and public policy discourse.[12]

Robert Greenwald, Larry Hedges, and Richard Laine (1996) pool findings from all published education production function studies that fit their criteria for inclusion.[13] Because research has not focused on teachers' scores, only 10 of the studies include measures of teacher test scores among the predictors of student achievement. Among these 10, most are over a decade old and use data from the 1960s and 1970s. The 10 studies include 24 independent coefficients measuring the relationship of teachers' scores to their students' standardized achievement scores. Among the 24 coefficients, 21 are positive, and only 3 are negative. Among the 21 positive coefficients, 12 are statistically significant at the .05 level. In their statistical meta-analysis, the authors address whether this pattern of coefficients across all of the studies might result purely by chance if there is no relationship in general between students' and teachers' scores. Their answer is unambiguously no. Some aspects of the Greenwald, Hedges, and Laine analysis have been challenged on methodological grounds, but the findings regarding teacher test scores have not; other methods of summarizing the literature would lead to the same general conclusion.[14]

Simply stated, even though the number of studies is relatively small, it appears generally that teachers who score higher on tests produce students who do also. Let me emphasize that no one characteristic of a teacher is a

[12] For example, see the response by Cizek (1995) to King (1993) in the *Review of Educational Research.*

[13] These standards were (1) the study was in a scholarly publication (e.g., a refereed journal or a book); (2) the data were from schools in the United States; (3) the outcome measure was some form of academic achievement; (4) the level of aggregation was at the level of the school district or a smaller unit; (5) the model controlled for socioeconomic characteristics or for prior performance levels; and (6) each equation was stochastically independent of others, such that only one of several equations from a study that used the same students but different outcome measures (e.g., math scores in one equation but reading scores in another) was kept for the Greenwald, Hedges, and Laine analysis.

[14] A response by Hanushek (1996) disputes the Greenwald, Hedges, and Laine interpretation of the evidence regarding school resources, especially class size. Hanushek does not dispute the findings regarding teachers' test scores, however. This is apparently because the vote-counting method that Hanushek tends to prefer would produce the same basic conclusion about teachers' scores. There is no clear winner of the debate regarding the methodological issues, because each side is correct if its favored assumptions are true, and there is no neutral way to test the validity of the assumptions.

totally reliable predictor of his or her performance. Nor are most teachers uniformly strong or weak in every subject or with all types of students. Nevertheless, until we develop more and better research to test it more completely and rigorously, my judgment is that a positive causal relationship between students' and teachers' scores should be the working assumption among policymakers.[15] Below, I present some more detailed evidence that supports this judgement.

Evidence from Texas and Alabama[16]

During the late 1980s, I constructed a data set for about 900 districts in Texas. Data were aggregated to the district level, and variables included teachers' scores on the Texas Examination of Current Administrators and Teachers (TECAT). Texas required all of its teachers to pass the TECAT or relinquish their jobs in 1986.[17] The test was essentially a reading, vocabulary, and language skills test, geared to about an eleventh grade level of difficulty. Some teachers had to retake it, but most eventually passed. Controlling statistically for a host of school and community characteristics, I found that district-average TECAT scores were strong predictors of why some school districts had higher student reading and math scores and larger year-to-year gains (Ferguson 1991). Thus, at least for Texas, a certification test that measured no specific teaching skills and that challenged teachers at only an eleventh grade level of difficulty seemed to distinguish among levels of teacher effectiveness. Later, Helen Ladd and I found similar patterns for Alabama, using teachers' college entrance exam (i.e., ACT) scores from when they applied to college.[18]

[15] One hypothesis is that teachers who score high on tests are good at teaching students to do well on tests or that they place greater emphasis on test taking skills, and that is why their students score higher. By this hypothesis, test score differences overstate "true" differences in how much children have learned. I have found no research that tries to test the validity of this hypothesis or to gauge the magnitude of any associated overstatement of differences in learning.

[16] This section draws extensively from an earlier paper of mine (Ferguson 1998) with the permission of Brookings Institution Press.

[17] Not many lost their jobs in Texas because with second and third chances most passed (I do not know the details for Arkansas). For the story of what happened in Texas, see Shepard and Kreitzer (1987).

[18] See Ferguson and Ladd (1996).

Below I present some new estimates, using the Texas and Alabama data and adding a few distinctions to my previous work.[19] I review evidence from Texas in the 1980s showing that teachers' scores were lower where larger percentages of students were black or Hispanic.[20] I also present evidence that test score gaps among teachers contributed to the black-white test score gap among students. Further, I use data from Alabama to show that when certification testing reduces entry into teaching by people with weak basic skills, it narrows the skill gap between new black and white teachers. Finally, I suggest that because rejected candidates would probably have taught disproportionately in black districts, initial certification testing for teachers is probably helping to narrow the test score gap between black and white students in Alabama.

Teachers' Scores and Students' Race-Ethnicity

Texas tested all of its teachers in 1986 using the TECAT. Black teachers had lower scores than white teachers by more than a standard deviation, and black teachers were more likely than white teachers to teach in districts with many black students.[21] (See column 1 of table 1.) Moreover, white teachers who taught in more heavily black and Hispanic districts tended to have lower scores than other white teachers. (See column 2 of table 1.) In Texas, and certainly in other places too, attracting and retaining talented people with strong skills to teach in the districts where black and Hispanic students are heavily represented is part of the unfinished business of equalizing educational opportunity.

[19] Specifically, the earlier paper (Ferguson 1991) did not have separate scores for elementary and high school teachers. Now, having both elementary and high school teachers' scores provides the basis for testing whether the difference between third-to-fifth and ninth-to-eleventh grade gains is a function of the difference between elementary and high school teachers' scores. See below.

[20] For more statistical estimates using these data, see Ferguson (1991). Kain is also currently assembling a large data set for Texas with which to study student performance at the individual level. See Kain (1995) and Kain and Singleton (1996) for two early papers from the project.

[21] This standard deviation is for the statewide distribution of scores among individual teachers.

Table 1. Effects of Percent Minority on Teachers' Test Scores and of Teacher Test Scores on Student Mathematics Achievement, Texas (District-level data; *t*-statistics in parentheses)

Dependent Variables:							
	Teachers' TECAT Scores		Math Scores for Students in 1988				
Explanatory variable	Black Teachers	White Teachers	Fifth Grade		Eleventh Grade		Difference in Gains: 9th–11th minus 3rd–5th
Column	1	2	3	4	5	6	7
TECAT Scores							
HS Teachers	...	...	...	...	...	0.128 (3.83)	0.164 (2.09)
Elem. Teachers	...	...	...	0.146 (2.96)	...	...	-0.179 (2.13)
Percent Minority Students:							
Black	-0.031 (3.52)	-0.014 (9.58)	-0.006 (1.81)	-0.0004 (0.12)	-0.019 (7.83)	-0.015 (5.70)	-0.012 (2.07)
Hispanic	-0.013 (2.93)	-0.010 (14.03)	-0.008 (3.35)	-0.007 (2.78)	-0.014 (8.33)	-0.013 (7.73)	-0.007 (1.91)
Mathematics scores, 1986							
Third Grade	...	...	0.394 (12.08)	0.367 (11.41)	...	...	...
Fifth Grade	...	...	...	...	0.455 (16.92)	0.438 (16.15)	...

Table 1. Effects of Percent Minority on Teachers' Test Scores and of Teacher Test Scores on Student Mathematics Achievement, Texas (District-level data; *t*-statistics in parentheses) (continued)

Additional Var's*	a	a	b	b	b	b	b
N	386	919	884	884	853	853	849
Adj. R-squared	0.03	0.19	0.46	0.47	0.72	0.73	0.04

NOTE: Observations are for districts, and data are district-level averages. Teachers' TECAT scores and students' math scores are measured in standard deviations of the mean scores among districts. For both TECAT scores and students' scores, this standard deviation from the distribution of district-level means is roughly one-third the standard deviation of the statewide distribution of scores among individuals. Each observation in each regression is weighted by the square root of student enrollment.

*Additional Variables: (a) Constant term is only control variable. (b) Control variables: teachers per student; percent of teachers with master's degrees; percent of teachers with 5 or more years experience; percent of adult population with more than high school education; percent of adult population with high school education; log per capita income; percent of children in poverty; percent of female-headed families; percent in public schools; percent of migrant farmworker students; percent English-as-a-second language; and indicator variables for city, suburb, town, nonmetro city, rural, Mexican border high-poverty district.

Teachers' Scores Help Predict Racial Test Score Gaps

Estimates using the Texas data and standard econometric specifications for education production functions show that TECAT scores are important predictors of students' math scores. (See columns 4 and 6 of table 1.)[22] In addition, teachers' scores help to explain why average math scores are lower in districts where larger percentages of students are black.[23] However, we cannot be sure that teachers' test scores affect students' test scores, because teachers' scores might merely be standing in for some omitted variables that are correlated with both teachers' and students' scores. Fortunately, separate scores for elementary and high school teachers allow me to circumvent this problem.[24] I compare high school gains to elementary school gains in the same district and ask whether the difference in high school and elementary school gains is larger in districts where the TECAT gap between high school and elementary school teachers is larger.[25] Using this approach, a change of one standard deviation in teachers' TECAT scores predicts a change of 0.17 standard deviation in students' scores over the course of two years.[26]

[22] Table 1 shows regression results where the dependent variable is the math score in 1988 for fifth grade (columns 3 and 4) or eleventh grade (columns 5 and 6). Two of the four columns include teachers' scores among the explanatory variables. All four columns include math scores for the same cohort from 1986. Including earlier scores for the same cohort among the explanatory variables is a standard way of estimating gains in achievement since the earlier date.

All of the regressions reported in table 1 are weighted by the square roots of district enrollment. This is a standard fix-up for heteroskedasticity in cases where data are means from samples of different sizes. Houston and Dallas are not included in the analysis because of a poorly conceived decision that I made when constructing the data set several years ago. For a detailed description of the data, see Ferguson (1991).

[23] Compare the coefficient on "percent black among students" from column 3 with that in column 4; and compare the coefficient in column 5 with that in column 6. Note that percents black and Hispanic are on a scale of 0 to 100.

[24] This of course assumes that unmeasured factors affecting *differences* between elementary and secondary students' test score gains are not correlated positively with *differences* between elementary and secondary teachers' scores.

[25] The dependent variable in column 7 of table 8 is the difference between two differences: (a) the district's mean high-school gain between the ninth and the eleventh grades, minus (b) the district's mean math score gain between the third and the fifth grades. Elementary and high school teachers' TECAT scores are included as separate variables.

[26] Here, 0.17 is the average of 0.164 (the coefficient on high school teachers' scores) and 0.179 (the absolute value of the coefficient on elementary school teachers' scores) from column 7 of table 8.

If the impact of skilled teachers is important *and accumulates*, then unusually high (or low) average TECAT scores for an entire district should help to pull up (or down) students' scores, and this impact should become more starkly apparent, the longer children are in school. For example, among districts where students do poorly in the early years of elementary school, districts where TECAT scores are unusually high should achieve much higher student scores by the end of high school than districts where TECAT scores are unusually low. To test this, I selected four sets of districts for comparison: districts with unusually high TECAT scores but low first- and third grade math scores (N=3); districts with unusually high TECAT scores and high first- and third grade math scores (N=37); districts with unusually low TECAT scores and low first- and third grade math scores (N=25); and districts with unusually low TECAT scores and high first- and third grade math scores (N=4).[27]

For each of the four sets of districts, figure 1 graphs the district-average math score for grades 1, 3, 5, 7, 9 and 11 for the 1985–86 school year.[28] Compare the patterns for districts that have similar teachers' scores. The dashed lines are districts where teachers' scores are more than a standard deviation above the statewide mean. Even though they start at opposite extremes for first- and third grade scores, the two have converged completely by the 11th grade. The solid lines are districts where teachers' scores are more than a standard deviation below the statewide mean. Here too, students' scores have converged by the eleventh grade, but at a far lower level.

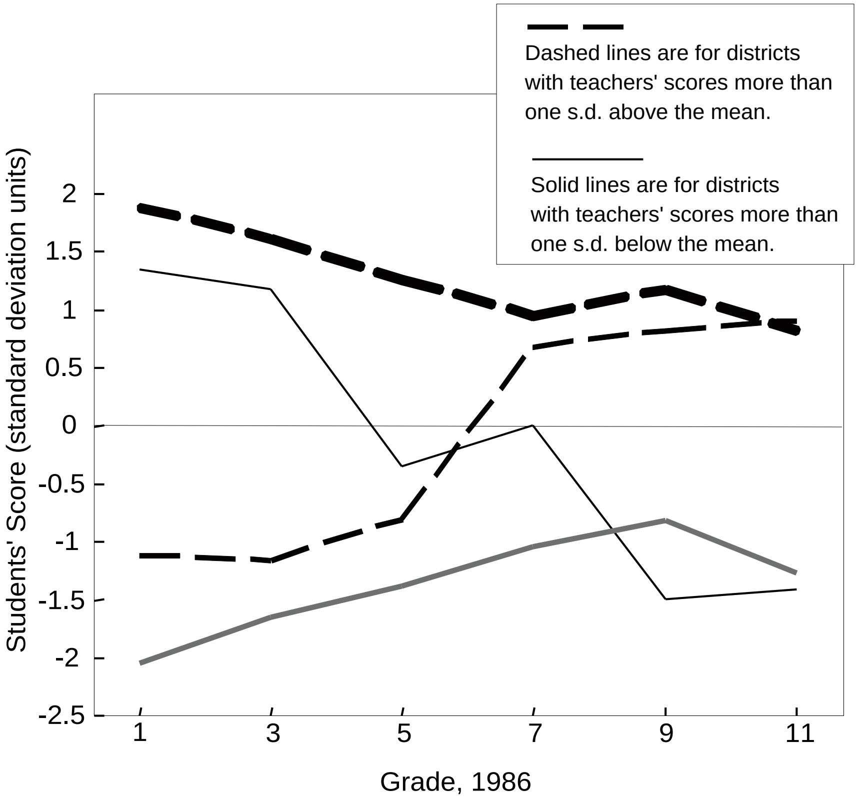
Figure 1 is not absolute proof of causation, but it is exactly what one would expect under the assumption that teachers' measured skills are important determinants of students' scores. Also, the magnitude of the change in

[27] I define "unusually" high (or low) to be a district-average TECAT score of more than one standard deviation above (or below) the statewide mean, where the relevant standard deviation is that among district-level means. Districts with low first and third grade math scores are those where math scores are more than a half standard deviation below the statewide mean for both years. Here too, the relevant standard deviation is that among district-level means. For both students' and teachers' scores, the ratio of statewide individual-level to district-level standard deviations in these data is 3 to 1.

Districts with high-scoring teachers and low-scoring students or low-scoring teachers and high-scoring students are rare. This is why, from roughly 900 districts, I could identify only a few, as indicated in the text and in the note to figure 1.

[28] A diagram for students' reading scores (not shown) follows the same general pattern, as do similar graphs using data for Alabama, albeit less dramatically.

Figure 1. Effect of Teachers' Test Scores on District-Average Mathematics Test Scores across Grades, Texas, Selected Districts, 1985–86



SOURCE: Author's calculations based on data obtained from the Texas Education Agency.

NOTE: Sample comprises three districts with unusually high teacher scores on the Texas Examination of Current Administrators and Teachers and unusually low scores on first and third grade mathematics achievement tests; four districts with low teacher scores and high first and third grade student scores; 37 districts with high scores for both teachers and students; and 25 districts with low scores for both teachers and students. For TECAT scores, "high" and "low" mean one standard deviation or more above and below, respectively, the Texas mean; for mathematics scores, the respective criteria are 0.50 standard deviations above and below the Texas mean. Standard deviations for both teachers' and students' scores are from the distribution of district-level means. In each case, the ratio of this standard deviation to that for individuals statewide is 3 to 1.

figure 1 from elementary through high school is almost exactly what one would predict using the regression estimates from column 7 of table 1. Specifically, for two districts starting with equal student scores, but teachers' scores separated by two standard deviations, over 10 years the difference in student scores would accumulate to 1.70 standard deviations.[29] This is a large effect.[30]

Certification Testing Probably Narrows the Black-White Test Score Gap

Relying less on evidence from research than on their own judgment, policymakers in 43 states had enacted some form of initial competency testing for teachers as of 1996.[31] Thirty-nine states include a test of basic reading and (sometimes) math skills. This is usually supplemented by an additional test of professional knowledge, such as the National Teachers Exam (NTE, now called PRAXIS), which is (as of 1996) used in 21 states.

Initial certification testing restricts entry into the teaching profession. Figure 2 shows the effect of certification testing on the mix of people who became teachers after Alabama began requiring certification tests in 1981. The data are from teachers' ACT scores at the time they applied to college.[32] After certification testing began, the test score gap between new black and white teachers fell sharply. Since districts in Alabama that have more black students also have more black teachers,[33] a change that increases the average level of skill among incoming black teachers should disproportionately benefit black children. If this pattern recurs in other states, as seems likely, we should find that black children's scores improve more than white children's scores after states

[29] 1.70=0.17 x 2 s.d. x 5 two-year intervals.

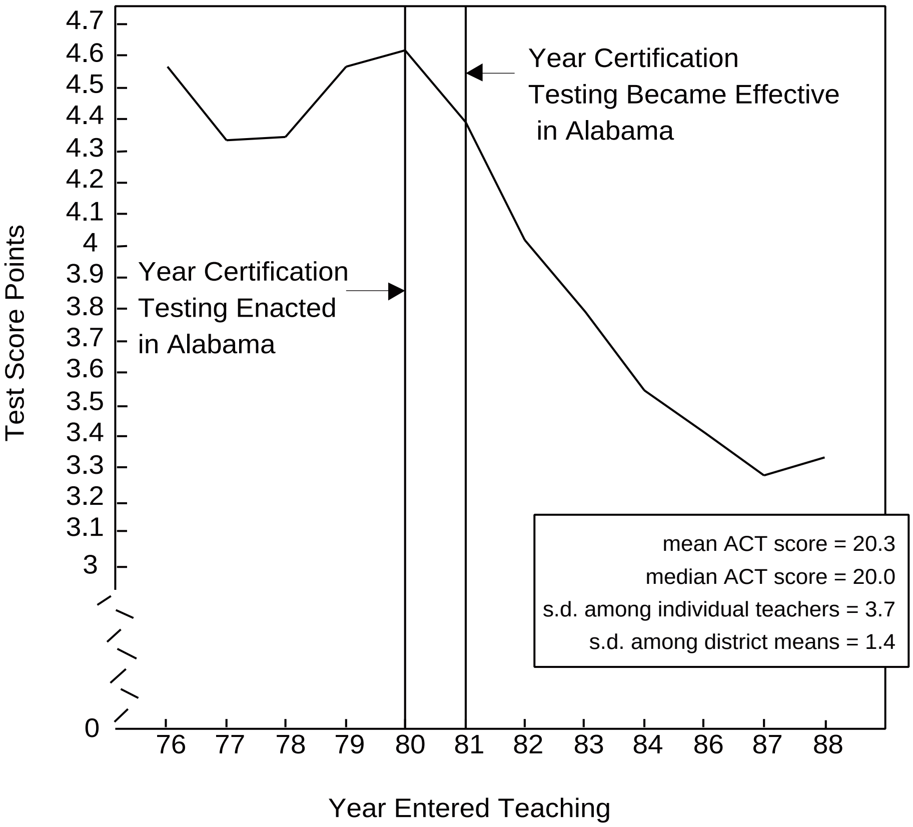
[30] This is not simply regression to the mean for student scores. Note that there are *two* sets of districts whose student scores are far below the mean as of the first and third grades. Only the districts with high teacher scores have student scores above the mean by the end of high school. Scores do regress toward the mean for the districts with low teacher scores, but these student scores nevertheless remain substantially below the mean. A similar set of statements applies to the districts whose first and third grade scores are above the mean.

[31] See U.S. Department of Education (1996). Table 154.

[32] See Ferguson and Ladd (1996) for more detail on the ACT data for Alabama.

[33] The simple correlation of "percent black among students" and "percent black among teachers" is 0.91 among 129 districts in Alabama.

Figure 2. Difference between Mean College Entrance Exam Scores of White and Black Teachers by Year of Entry into the Profession, Alabama, 1976–88



SOURCE: Author's calculations, unpublished data. ACT scores are from teachers' college entrance exams and are not associated with any certification exams that they may have taken.

implement certification testing for teachers (but we should expect some improvement even for whites).

Twenty-five years ago, working with data from the 1966 Coleman report, David Armor wrote:

> Even though black teachers' formal training seems as extensive as that of white teachers, if not more so, their verbal scores indicate that they have far less academic achievement. It is especially

> ironic, when schools are concerned with raising black student achievement, that the black teachers who have the major responsibility for it suffer from the same disadvantage as their students.[34]

Once certification testing began in earnest after 1980, passing rates for black applicants in states across the nation were sometimes half those for whites.[35] Certainly, some black teachers who failed would have become good teachers. However, the relevant policy question is whether students on average are better off with the policy in place. I think the answer is yes.[36] However, truly definitive answers would require better data, developed and utilized in a multistate, longitudinal program of research.

We Need Better Data

Testing whether teachers' test scores or observers' ratings are good predictors of professional effectiveness is not a simple process. Even when there is agreement that gains in pupils' test scores should be the primary measure of professional output, a number of statistical assumptions must hold in order for studies to produce reliable estimates of how well teachers' scores (or ratings) measure their effectiveness. Problems associated with measurement error, incorrect functional forms, omitted variable bias, simultaneity bias, and reverse causation plague this type of analysis. Authors in the education production function literature over the last few decades have encountered these problems routinely (but seldom overcome them). In addition, during the 1990s, statisticians have emphasized the importance of hierarchical models to distinguish student-level from school-level from district-level effects of explanatory variables.[37]

[34] See Armor (1972).

[35] Quoting numbers from Anrig (1986), Irvine (1990, p. 39) presents the following numbers: "In California, the passing rate for white test-takers was 76 percent, but 26 percent for blacks; in Georgia, 87 percent of whites passed the test on the first try, while only 34 percent of blacks did; in Oklahoma, there was a 79 percent pass rate for whites and 48 percent for blacks; in Florida, an 83 percent pass rate for whites, 35 percent for blacks; in Louisiana, 78 percent for whites, 15 percent for blacks; on the NTE Core Battery, 94 percent of whites passed, compared with 48 percent of blacks."

[36] Available estimates suggest that the impact of teachers' scores on students' scores does not depend on the race of the teacher. Ehrenberg and Brewer find this using the verbal skills test from Coleman (1966). I also find it in unpublished results using data for Texas.

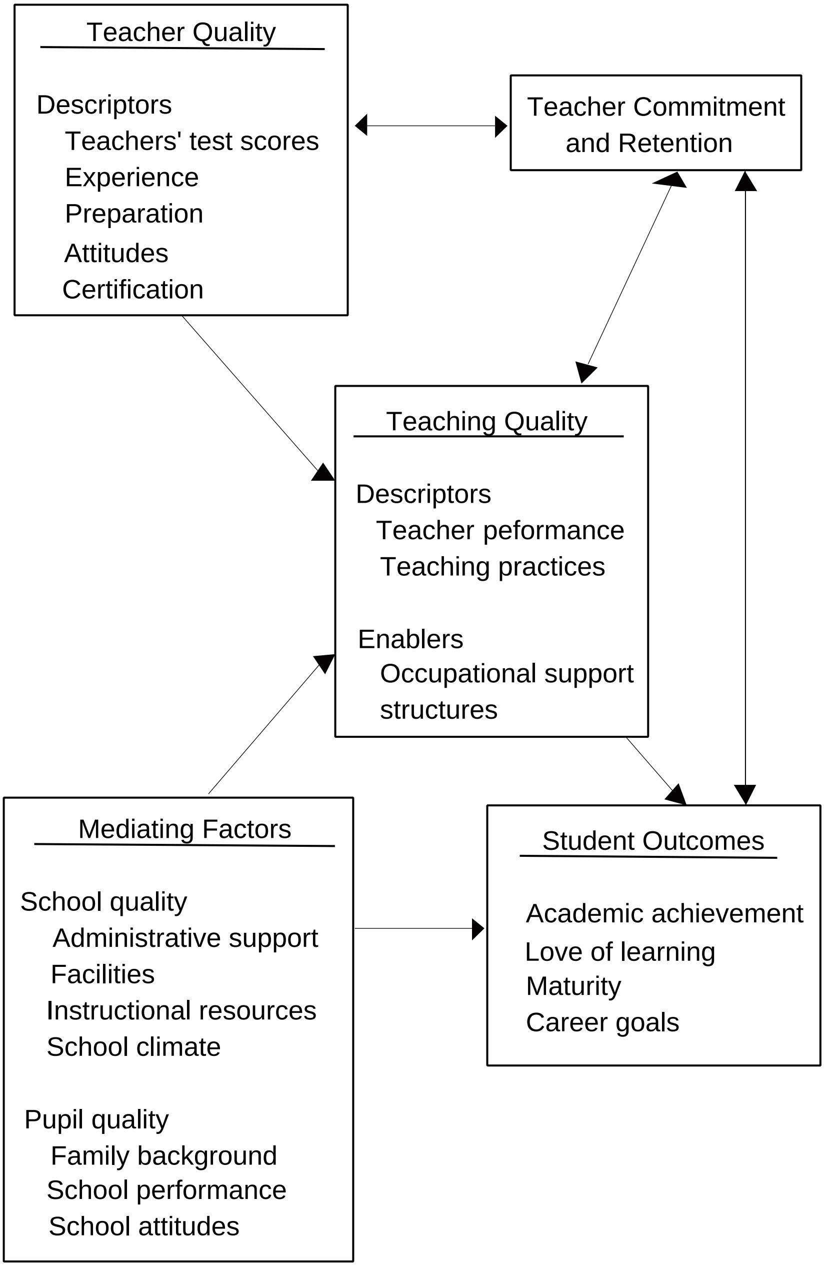
[37] Other papers in this volume elaborate the advantages of multilevel modeling.

Teacher quality measures present all of the standard statistical problems listed in the paragraph above: teacher quality, including teachers' test scores, are measured with error; student quality in a school or district can affect which teachers choose to apply there, creating reverse causation from student performance to teacher quality; particular measures of teacher quality may matter more or less depending on other variables, such as class size, so simple linear models that ignore interactions may produce misleading results; correlations between teacher quality and other inputs such as parental effectiveness can produce biased estimates for the effect of teacher quality if parental variables are omitted from the analysis or measured with considerable error. Further, most studies lack the type of data necessary for sorting out the issues that the advocates of multilevel estimation emphasize.

Even when measuring the effect of teacher quality on student outcomes is the only goal, data requirements can be vast. It is difficult to emphasize enough that teaching is a complex process in which context matters. Helping students to achieve academic success, love of learning, maturity, or career success involves far more that high certification test scores. Indeed, Darling-Hammond and Hudson (1989) distinguish *teacher* quality (e.g., certification test scores, experience, preparation, attitudes, aptitudes) from *teaching* quality (i.e., performance in the classroom). Further, they point out that how effectively both teacher and teaching quality translate into student outcomes depends on characteristics of schools, students, and families. (See figure 3 for a summary picture.)

Since no analyst will ever achieve a fully specified statistical model of this process or have the ideal data for solving *all* the statistical problems listed above, we will never reach perfection. We can, however, do better than we have. Researchers seem to agree on at least two points. First, we need more random assignment experiments to test hypotheses about the productivity of schooling inputs such as small-versus-large classes or high-versus-low teacher test scores. Second, because random assignment studies are sometimes impractical, we need more student-level longitudinal data sets that include good measures of child, teacher, family, classroom, school, and community characteristics. Further, no matter how carefully we assemble longitudinal data, the possibility that results are driven, for example, by omitted variable bias, will always make the findings from individual studies less than definitive. Similarly, findings from a single random assignment study may depend on idiosyncratic conditions that are not maintained at other times and places. Hence,

Figure 3. A Model of Teacher-Quality Effects (Adapted from Darling-Hammond and Hudson, 1989)



for both random assignment experiments and statistical studies using longitudinal data, we need replication across multiple independent analyses. There has been no organized program of education research to test whether standard measures of teacher quality are reliable predictors of student learning.

Future Research Involving Teachers' Scores

The following are five sets of issues and questions that a future program of research could usefully address about measures of teacher quality, all involving teacher test scores.

1. **Best Practices**. Do teachers' own test scores predict whether they use practices in the classroom that researchers have classified as most effective?[38] Or is the apparent relationship between teachers' scores and student performance measuring something subtler than so-called best practices?[39]
2. **Fixed Effects**. Using fixed-effects specifications, researchers can estimate which teachers consistently over the years produce greater learning gains, as measured by changes in their students' standardized test scores.[40] Do teachers' own scores predict the teacher-effects that these studies estimate? If we put measures of effectiveness based on observer ratings into the same equations that include teachers' scores, does the predictive power of teachers' scores remain unshaken?
3. **Generalizability and Fairness**. Are teachers' scores equally accurate predictors for teachers with different characteristics (e.g., different ethnicities, different training, and so on)?
4. **Effects of Other Inputs**. Teachers' scores can also be control variables. More and better teacher test score data would provide better statistical controls for estimating the effects of other variables such as experience, masters degrees, class size, or even parents' education.

[38] For literature reviews regarding teaching quality and best practices, see Brophy (1986), Doyle (1986), Darling-Hammond and Hudson (1989), and Porter and Brophy (1988).

[39] For example, it could be that teachers with more skills are those with the better judgment — for example, those who *depart* from generally effective practices at precisely those times that the practices would not be effective.

[40] See Rivkin, Hanushek, and Kain (1998) and Sanders and Rivers (1996).

5. **Allocation of Teacher Quality**. Attracting more strong teaching candidates and having them teach where they are needed most is important. Who gets the best teachers and why? It would be useful to know the degree to which salaries and other factors are important predictors of where high-scoring teachers end up teaching (e.g., which grades, schools, tracks, districts).

Certainly, the list could be longer. However, a serious program of research that made important progress on these five sets of issues would be a big step forward.

The National Center for Education Statistics (NCES) could help in the following ways:

1. Information about teachers, for example, the college that the teacher attended for BA and MA degrees, could be added to the teacher surveys that accompany NCES student surveys.
2. NCES can convene and coordinate state-level researchers who are constructing longitudinal student-level data sets that include (or can include) teachers' scores and other teacher characteristics.
3. NCES can encourage the Educational Testing Service and other test makers to work with states to validate teacher exams as predictors of student performance.
4. NCES should increase the number of students sampled per teacher in longitudinal NCES data series. NCES could also facilitate matching of its data with state-level data for teachers and students.

None of these will be easy, but each would be helpful.

Conclusion

Difficulty talking in public about racial and ethnic differences in test score patterns is probably a major factor in why the nation has not addressed these issues with the seriousness that they deserve. This challenge needs to be confronted. As I write, public officials in the state of Massachusetts are debating whether to test incumbent teachers for recertification. The basis of their interest in testing is the belief that certain elements of core knowledge are foundations for professional practice. Any teacher who lacks this knowledge cannot, the theory goes, be an effective teacher. The bulk of the evidence that we have suggests that teachers' scores on even the most rudimentary of basic skills exams—for example, the 30-item test in the Coleman study or the TECAT test in Texas—can be statistically significant predictors of how much students will learn. Regarding whether to screen teacher candidates using such exams, I am inclined to give the benefit of the doubt to students, which for me means endorsing the continued use (*and ongoing improvement*) of certification exams. As Bernard Gifford suggested, it is better to work on raising the skills of teaching candidates who might otherwise fail than to lower the standards that teaching candidates are expected to meet, and thereby to raise the risk that children will receive poor schooling.

On the other hand, our knowledge is far from definitive and very incomplete. Current certification exams produce an unknown number of mistakes that cause individuals to suffer unfairly. Some candidates who rate high on dimensions that tests do not measure and who would have been good teachers fail certification exams and never become teachers. Conversely, some are "false positives" who pass the exams but may fail in the classroom. Nonwhite candidates are probably over-represented among the false negatives who fail the exams but would have been good teachers.[41] At the same time, nonwhite children are probably over-represented among beneficiaries. This is because more of the people who fail, and would *not* have been good teachers, would probably have shown up to teach in classrooms where nonwhite children are over-represented. We may never know for sure. Nonetheless, I believe that if we had better data, a greater willingness to debate hard questions, and a targeted program of research, we would find ways to be more nearly fair in selecting among teaching candidates and ultimately more effective in helping those hired to become good teachers.

[41] See the discussion in Jencks and Phillips (1998, 77). Assume that a test score is the only basis for selecting people into a job, such as teaching. Also assume that black candidates, on average, have lower average scores than whites but are more similar to whites on other skills that affect teaching quality. Jencks explains why a larger percentage of blacks will be excluded than whites, among those people who would have performed well if hired (or do perform well).

References

Anrig, G. R. (1986). Teacher Education and Teacher Training: The Rush to Mandate. *Phi Delta Kappan 67*: 447–451.

Armor, D. (1972). School and Family Effects on Black and White Achievement: A Re-examination of the USOE Data. In F. Mosteller and D. P. Moynihan (Eds.), *On Equality of Educational Opportunity.* New York: Random House.

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A. (1976, August). In E. Pauly and G. Zellman (Eds.) *Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools*. Report prepared by RAND for the Los Angeles Unified School District. Report Number R–2007–LAUSD.

Ballou, D., and Podgursky, M. (1997). *Teacher Pay and Teacher Quality*. Kalamazoo, MI: The W. E. Upjohn Institute for Employment Research.

Brophy, J. (1986). Teacher Influences on Student Achievement. *American Psychologist 41*(10): 1069–1077.

Cizek, G. J. (1995, spring). On the Limited Presence of African American Teachers: An Assessment of Research, Synthesis, and Policy Implications. *Review of Educational Research 65*(1): 78–92.

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.

Darling-Hammond, L. (1997). *Doing What Matters Most: Investing in Quality Teaching*. Kutztown, PA: National Commission on Teaching and America's Future.

Darling-Hammond, L., and Hudson, L. (1989). Teachers and Teaching. In R. J. Shavelson, L. M. McDonnell, and J. Oakes (Eds.), *Indicators for Monitoring Mathematics and Science Education.* Santa Monica, CA: RAND.

Doyle, W. (1986). Classroom Organization and Management. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching*, 3rd Edition. New York: MacMillan.

Ferguson, R. F. (1991, Summer). Paying for Public Education: New Evidence on How and Why Money Matters. *Harvard Journal on Legislation 28*(2): 465–498.

Ferguson, R. F. (1998). Can Schools Narrow the Black-White Test Score Gap? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington DC: Brookings Institution Press.

Ferguson, R. F., and Ladd, H. F. (1996). How and Why Money Matters: An Analysis of Alabama Schools. In H. F. Ladd (Ed.), *Holding Schools Accountable: Performance Based Reform in Education*. Washington, DC: Brookings Institution Press.

Gifford, B. R. (1985). Teacher Competency Testing and Its Effect on Minorities: Reflection and Recommendations. In E. E. Freeman (Ed.), *Educational Standards, Testing and Access: Proceedings of the 1984 ETS Forty-Fifth Invitational Conference.* Princeton, NJ: Educational Testing Service.

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

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

Hanushek, E. A. (1992). The Trade-Off between Child Quantity and Quality. *Journal of Political Economy 100*: 84–117.

Hanushek, E. A. (1996). A More Complete Picture of School Resource Policies. *Review of Educational Research 66*: 397–409.

Irvine, J. (1990). *Black Students and School Failure*. Westport, CN: Greenwood.

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

Kain, J. (1995). *Impact of Minority Suburbanization on the School Attendance and Achievement of Minority Children.* Final Report of Work Completed Under Spencer Foundation Small Grant. Harvard University, Department of Economics.

Kain, J. (1998, October). *The Impact of Individual Teachers and Peers on Individual Student Achievement.* Working paper. Cecil and Ida Green Center for the Study of Science and Society, University of Texas at Dallas.

Kain, J., and Singleton, K. (1996, May/June). Equality of Educational Opportunity Revisited. *New England Economic Review,* pp. 87–111.

King, S. H. (1993, summer). The Limited Presence of African-American Teachers. *Review of Educational Research 63*(2): 115–149.

Murnane, R. J., Singer, J. D., Willett, J. B., Kemple, J. J., and Olsen, R. J. (1991). *Who Will Teach? Policies That Matter*. Cambridge, MA: Harvard University Press.

Porter, A.C., and Brophy, J. (1988, May). Synthesis of Research on Good Teaching: Insights from the Work of the Institute for Research on Teaching. *Educational Leadership,* pp. 74–85.

Pressman, H., and Gartner, A. (1986, Summer). The New Racism in Education. *Social Policy 17*(1): 11–15.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (1998). *Teachers, Schools and Academic Achievement.* Paper presented at the January 1998 Annual Meeting of the American Economic Association, Chicago, IL.

Sanders, W. L., and Rivers, J. C. (1996). *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.* Research Report. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1998). The Tennessee Value-Added Assessment System (TVAAS): A Quantitative, Outcomes-Based Approach to Educational Assessment. In J. Milliman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, CA: Corwin Press.

Shepard, L. A., and Kreitzer, A. E. (1987, August-September). The Texas Teacher Test. *Educational Researcher,* pp. 22–31.

U.S. Department of Education. (1996). *Digest of Education Statistics*. National Center for Education Statistics.

Response: Two Studies of Academic Achievement

Robert M. Hauser
University of Wisconsin-Madison

The papers by Meredith Phillips (1998) and by Ronald F. Ferguson with Jordana Brown (1998) exemplify the best of contemporary educational policy research.[1] First, they focus on important questions: What are the sources of differentials in academic achievement between racial-ethnic groups in the United States? When do these differentials appear in the course of children's development? What is the role of family and school factors in the development of these differences? How can we best measure, understand, and reduce the differentials? How, if at all, do teacher qualification test scores—or other test scores—affect student learning? Should such test scores be used as a threshold for entry into the teaching profession? What are the effects of such tests on the qualifications of new entrants to teaching and on differentials in the test scores of teachers from majority and minority groups? Will smaller differences between the qualification test scores of majority and minority teachers lead to smaller differences in student achievement? What are the advantages and disadvantages of alternative measures of teacher quality?

Second, both papers use a wide array of evidence. Phillips focuses on new data from the Prospects study, but she—along with her collaborators in related work—actually draws on much of the accumulated evidence of trends and differentials in student achievement in the United States. Ferguson and Brown focus primarily on an important body of data on teacher test scores and student achievement for school districts in Texas, but they also draw on data from other states—notably Alabama—and from other recent studies of teacher qualifications and student performance. One need only think back to the mid-1960s, when the "Coleman-Campbell report" (1966) provided the *only* national

[1] The opinions expressed in this paper are those of the author. Address comments to Professor Robert M. Hauser, Department of Sociology, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, WI 53706.

data on educational resources and academic achievement, to realize that we have come some distance.

Third, both papers are methodologically and statistically sophisticated, with the authors arranging and examining evidence in new ways and, at the same time, not letting their work become model-driven to the point where they completely lost sight of the data or of the limits of their data in addressing their central questions. Indeed, both Phillips and Ferguson and Brown focus as much on better ways to ask their questions as on the important findings of their research.

Enough of generalities, what about the papers?

Response to the Phillips Paper

Phillips makes six main points. They are worth repeating, although I will quibble a bit with some of them.

The first point is as follows: “Traditional socioeconomic factors do not overlap with ethnicity as much as many people assume.” Ethnic differences in achievement are not easily reducible to socioeconomic or other social differences in academic achievement. Phillips observes that the reductionist view has been sustained in part by the political sensitivity of black-white differences. Thus, many researchers have tried to explain the gaps, as Phillips notes, by black-white differences in levels of family advantage, neighborhood poverty, or urban-suburban location. But these factors do not account for the test score gap.[2]

The second point is, “We should focus our surveys mainly on elementary school students rather than on high school students.” I think this is a bit overdrawn. To the degree that our focus is on academic achievement, the available evidence points to the malleability of learning in the early years. That is important. But we ought not to forget adolescence—recall the success of recent years in changing course content and requirements in high school—as well as the

[2] At the extreme, I have seen one leading economic scholar argue against adjusted statistical comparisons of educational outcomes between blacks and whites on the ground that there is not sufficient overlap of socioeconomic background to justify this form of comparison—a proposition that is patently contradicted by the evidence of overlap between distributions of social and economic standing in the black and white populations.

wider array of outcomes that determine what happens to youth when they leave high school (Hauser 1991). The downward drift of starting points of the major national longitudinal studies—from NLS 1972 to HS&B in the 1980s and NELS in the 1990s—has been a beneficial evolution. But we ought not to lose such samples as they age, no matter how young they are when we start. I will come back to this point again in discussing Phillips' fifth point.

The third point is as follows: "We should test children in both the fall and the spring of each school year." As Phillips notes, her evidence on this point from the Prospects study is compelling, and it builds on a decades-old history of similar findings. Why has this source of black-white test score differences not become a focus of public policy? What would it take to accomplish that? Need we wait until achievement test scores sink so low that the public approves test-based grade retention on a massive scale before we put any real money into summer school?

But I would question the calculus of Phillips' statistical comparisons of learning in summer school and during the school year. Such comparisons read *as if* score gains during the school year are the work of schools alone, while summer gains or losses are the work of families alone. Consider alternative assumptions: Suppose learning is linear in exposure to learning environments. Students do not leave their families during the academic year; they spend more time in school and somewhat less with families. Suppose we ignore summer school, and attribute summer gains or losses to families. Then, summer changes reflect the effects of families (including peers and neighbors), while changes during the academic year reflect the combined effects of schools and families. Assume, further, that exposure to school and family is equal throughout the academic year. Now, for example, look at the top row of table 6. Three months of family-only exposure in the summer produces a black loss of 20.47 points. This implies a loss of 20.47/6 = 3.41 points per month during the school year, assuming summer is 3 months long and family exposure is half as great during the academic year as in the summer. The implied loss is 3.41 x 9 = 30.71 points during the academic year. Since black children gain relative to whites during the academic year—by 9.27 points—the implication is that the annual effect of schooling is 9.27 + 30.71 = 39.98 points. In this account, schooling plays an enormously effective role in reducing black-white test score differences.

Please do not take this account too seriously. In particular, the test used in the Prospects study is vertically equated to show larger gains at low than at high performance levels. Thus, the score gains of African American students are not strictly comparable to those of majority students. My assumptions and calculations are no more than illustrative. But they are, I think, worth thinking about. What is the role of schools in learning relative to families during the school year? During the summer? How could we learn more about it? Do family effects really offset school effects, or are they complementary? If so, how do we explain the summer deficits? What would be the long-term benefits of year-round schooling, and how could we realize them?

Phillips' fourth point is, "Tests of seemingly similar skills … sometimes yield very different estimates of ethnic differences in achievement." Here, the evidence provided by Phillips (in table 7) appears supportive, but I am not sure that it is strong enough. The problem is that the measures of math concepts and math computations are not independent, so simple comparisons of means and their reported errors are not appropriate to test differences in the effects of ethnicity on the outcomes. A bit more modeling is required.

The fifth point cited is as follows: "Different surveys of apparently similar populations sometimes yield contradictory results." I am not at all convinced by the comparison of children of the NLSY with those of Prospects in table 8. The key issue here is "apparently similar populations." The CNLSY is a household-based survey, and children of women in the NLSY of 1979 have passed through school over a period of years, assuming that Phillips has captured the experience of those children in full. Those children do not represent all children in the birth cohorts because children in the same years may be born to mothers outside the cohorts of the NLSY. Children of the NLSY are subject to attrition from both the parent and child samples. Children of the NLSY do not include children of recent immigrants from the same or different cohorts as the mothers of the NLSY. I am not at all sure that it is worth trying to reconcile all of the differences between Prospects and CNLSY; I am reasonably sure that the fact that the surveys yield discrepant findings does not in itself justify a call for multiple, independent survey operations.

The sixth point is, "The vocabulary gap between African Americans and European Americans is already large by the time children are three years old." I agree, and that's why we are here.

What survey research designs might address Phillips' concerns? I would suggest, and not for the first time, that the need for replicate observations and for alternative methods should be met by the regular initiation of new, and perhaps modestly sized, longitudinal cohort surveys—and not by larger, one-time-only or once-per-decade surveys. I have made the same proposal for studies of adolescent development. 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 opportunity for core content stable enough to permit aggregation of findings across cohorts to yield greater statistical power. There is already a considerable literature on the need for such surveys and on possible designs (National Research Council 1995). We need not reinvent it here.

Response to the Ferguson and Brown Paper

Ferguson and Brown (1998) focus primarily on the effects of teacher test scores on achievement test scores in Texas. They briefly consider other measures of teacher effectiveness: classroom observation—which they dismiss as too costly and of doubtful validity—and direct observation of student gains in test scores. They dismiss the latter as requiring years of observed data but note, "When the data are adequate and the analysis is done appropriately, [this is] probably the best that we can do at measuring the effectiveness of individual teachers." I agree about the validity of this method and wonder why it is not viewed as more practical for the evaluation of teachers beyond point-of-hire. Many of us—as college and university academics—have had judgments made about our effectiveness and competence on the basis of accumulated dossiers. To be sure, these are lists of books, papers, and talks, rather than raw scores, but the principle is the same.

Ferguson and Brown's analyses of the Texas data are for school districts as units. Since the ratio of student test score standard deviations at the district level to those at the individual level is as 1 to 3—and similarly for teacher test scores—much of the analysis of statistical findings passes transparently from one level to the other. This is convenient, but perhaps too much so. Districts are not students, and the specification problems that seem obvious to us when we think about determinants of individual test scores—many of which have been satisfied by Ferguson and Brown—may not be the right ones to solve at their level of aggregation. They know it, and they say it, but it remains a puzzle. The 1 to 3 ratio of standard deviations has an important implication that goes unstated in the text: 90 percent of the variance in student test scores, like 90 percent of the variance in teacher test scores, lies within districts. There is a lot of room for specification error at the latter level to escape our notice. Think, for example, of the issue of validity of certification tests such as the TECAT, and ask yourself whether the Texas school district data bear on that issue.

I also worry about the fact that the data are from 1985–86, just at the time teacher testing was introduced. What has happened to the distributions of teacher test scores, both within and between districts, since that time? Are the standard deviations of 1985–86 the right metric for us to use in thinking about policies in the late 1990s?

Ferguson and Brown are both clever and wise in their statistical analysis. I particularly commend the use of a “difference of differences” estimator of the effect of teacher test scores on student gains, reported in table 1. Similarly, I like the fact that they help us look directly at the data in figure 1, which is one of the most fascinating statistical graphics I have seen in some time. It is a striking example of what demographers would call a synthetic cohort analysis, in which variation of test scores across grade levels within a single year is taken as a proxy for variation in achievement within a single cohort across its progression through grade levels. What the figure appears to show is that academic achievement tends to become consistent with initial teacher scores in the cases where substantial inconsistencies occur in the lower grades. There are some reasons to be wary of this finding: The number of inconsistent districts is very small, and we do not know why they are inconsistent. And the data pertain to synthetic cohorts, not real changes in academic achievement across time.

Finally, if I understand the graphic correctly, this is one case where the distinction between measurements in standard deviations at the district versus individual levels really makes a difference. Imagine rescaling figure 1 in standard deviation units of individual test scores. In this case, if I follow the arithmetic, a consistent, 10-year improvement of teacher scores by two standard deviations—how feasible is such a gain?—would accumulate to 0.57 standard deviations. That is a substantial fraction of initial black-white differences in test scores. Is the evidence strong enough to support such a conclusion?

I would add that this is a striking, but perhaps too limited, example of our need to get closer to the data. We should be doing a great deal more exploratory data analysis, even in situations where we think we know how to model data successfully. Even in large, longitudinal surveys, that may help us as much in putting together a coherent story as any number of smaller, even smaller and richer, studies. Model the data, but also look at the data.

One other question about the Ferguson-Brown paper strikes me as particularly important. It is mentioned at the close of the paper. How much of the measurable difference in teacher effectiveness can be attributed to test score differences? That is, suppose we ran the dummy variable regressions of student test score change on “teacher” as described at the beginning of the paper. What would happen to the coefficients of teachers as their test scores enter the equation? And what other teacher characteristics would explain the remaining effects? We might ask, also, whether the effects of teachers’ test scores are diagnostic or causal. That is, do they truly account for teachers’ effectiveness, or are they merely sound evidence to be used in screening potential teachers? One way or the other, what are the costs and benefits of improved supervision and training relative to—or complementary to—the skills and knowledge that teachers initially bring to the job? For example, what should we make of the evidence that the support of teaching and teachers is a major impediment to the success of standards-based reform?

Persistent Issues in Educational Policy Research

These two fine papers also remind me of potential weaknesses and points of contention in contemporary educational policy research. Two of these points of contention are the centrality of test scores as educational outcomes and a possible failure to respect the limits of observational data in answering policy questions that can only be answered in the language of cause and effect. This is not news, but I think the main points bear repeating.

We are here to think about academic achievement—how it is produced, how it becomes differentiated, how to measure it, how to measure its production. This is all well and good: I do not want to be one of those miserable critics who say that we should not be here doing what we are doing today. Student learning is a main objective of schooling, and achievement tests are a great social invention as well as our main way of measuring student learning. But we in the research and policy community can focus too much on tests and testing. The focus on student learning as a key outcome of schooling devolves into a focus on student test scores as a key outcome of schooling, and the latter may devolve into a focus on student test scores as the *only* outcome of schooling. To paraphrase Vince Lombardi's remark concerning winning, "Test scores are not the main thing; they are the only thing." As researchers, we may learn all about academic achievement—and little else. As a nation, we may get what we wish for and live to regret it.[3]

If we know all too little about educational production functions, we should be even more humble about our understanding of what makes people healthy, wealthy, and wise. For 30 years I have been watching as the 10,000 students in the Wisconsin Longitudinal Study have marched through life. This is the same cohort of 1957 high school graduates portrayed in the situation comedy, *Happy Days*. I have learned two things as I (along with my colleagues) have watched trajectories of schooling, jobs, and family lives, and of states of depression and well-being and of health and disease. The first—to use a quip by Paul Siegel from some years ago—is that everything that happens to you before your sixteenth birthday affects everything that happens to you after your sixteenth birthday by way of the amount of schooling that you finish. The second is that adolescent test scores provide no exception to the rule.[4] Education is not just test scores, and we should not wish to make it so. Education is a fascinating bundle of learning and motivation, of values and skills, of behaviors and—yes—certification, and the easy part of our job is to unbundle it. The hard part is not to lose sight of the whole. In a smaller, but older longitudinal study, the late social psychologist, John Clausen (1991, 1993) summarized the key to the good life as *planful competence*—a combination of academic success with responsibility and motivation.

All of this broadens the subject without any reference to the demography of schooling, with which the connections with academic achievement are pervasive, complex—and largely ignored. To go back to the problem of getting what we wish for, I think it is fair to say that we in the research policy community have aroused, and now bear responsibility for moderating, the national mania for achievement testing. Read *High Stakes: Testing for Tracking, Pro-*

[3] I am reminded of the urban renewal program of the late 1950s and early 1960s, in which the goal was "decent, safe, and sanitary housing." That is just what we got, but only for a short time, and what we did not get was healthy, viable communities.

[4] For example, see Hauser and Sweeney (1997).

motion, and Graduation, the report of the National Research Council (1999), if you want to learn more about both of these last two points.

In the closing passage of her paper, Meredith Phillips rightly observes the distinction between *understanding* the growth of differentials in academic achievement—the main focus of her work—and *changing* those differentials—a task for which, she argues, we should turn to large-scale field experiments. Similarly, Ferguson and Brown muse about the limits of econometric methodology in explicating the role of teacher qualifications in student achievement. I would put the matter somewhat differently, i.e., observational studies, even those designed and carried out to the highest standards, are mainly useful in telling us what has happened, when, and to whom. I am all in favor of putting such accounts into the form of statistical models, to the extent justified by the data and by prior knowledge and plausible assumption. Such exercises are most valuable—witness Meredith Phillips' compelling finding that summer deficits dominate winter surpluses of learning among black schoolchildren. But they do not tell us "how to fix it." The language of causality provides a useful way of thinking about the world, but we ought not to invest it with more belief than our research designs and evidence can sustain.[5]

[5] On the other hand, research on experimental and nonexperimental methods of evaluating welfare policies and reform provides equally cautionary evidence about the value of observational data—especially when we take the final leap between theory and practice.

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.

Clausen, J.A. (1991). Adolescent Competence and the Shaping of the Life Course. *American Journal of Sociology 96*(4): 805–842.

Clausen, J.A. (1993). *American Lives: Looking Back at the Children of the Great Depression.* New York: The Free Press.

Hauser, R.M. (1991, fall-winter). What Happens to Youth After High School? *IRP FOCUS 13:* 1–13.

Hauser, R.M., and Sweeney, M.M. (1997). Does Adolescent Poverty Affect the Life Chances of High School Graduates? In G. Duncan and J. Brooks-Gunn (Eds.), *Growing Up Poor* (pp. 541–595). New York: Russell Sage Foundation.

Herrnstein, R.J., and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: The Free Press.

National Research Council. (1995). *Integrating Federal Statistics on Children.* Washington, DC: National Academy Press.

National Research Council. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation.* In J.P. Heubert and R.M. Hauser (Eds.), Washington, DC: National Academy Press.