EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Sherman Dorn
College of Education
University of South Florida

Volume 14 Number 8 March 20, 2006 ISSN 1068–2341

# The Black-White Achievement Gap: Do State Policies Matter?

Henry I. Braun
Educational Testing Service

Aubrey Wang
School District of Philadelphia

Frank Jenkins
Westat, Inc.

Elliot Weinbaum
University of Pennsylvania

Citation: Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black-White achievement gap: Do state policies matter? *Education Policy Analysis Archives, 14*(8). Retrieved [date] from http://epaa.asu.edu/epaa/v14n8/.

## Abstract

A longstanding issue in American education is the gap in academic achievement between majority and minority students. The goal of this study is to accumulate and evaluate evidence on the relationship between state education policies and changes in the Black-White achievement gap, while addressing some of the methodological issues that have led to differences in interpretations of earlier findings. To that end, we consider the experiences of ten states that together enroll more than forty percent of the nation's Black students. We estimate the trajectories of Black student and White student achievement on the NAEP 8th grade mathematics assessment over the period 1992 to 2000, and examine the achievement gap at three levels of aggregation: the state as a whole, groups of schools (strata) within a state defined by the SES level of the student population, and within schools within a stratum within a state. From 1992 to 2000, at every level of aggregation, mean achievement rose for both Black students and White students. However, for most states the achievement gaps were large and changed very little at every level of aggregation. The gaps are pervasive, profound and persistent.

There is substantial heterogeneity among states in the types of policies they pursued, as well as the coherence and consistency of those policies during the period 1988–1998. We find that states' overall policy rankings (based on our review of the data) correlate moderately with their record in improving Black student achievement but are somewhat less useful in predicting their record with respect to reducing the achievement gaps. States' rankings on commitment to teacher quality correlate almost as well as did the overall policy ranking. Thus, state reform efforts are a blunt tool, but a tool nonetheless.

Our findings are consistent with the following recommendations: states' reform efforts should be built on broad-based support and buffered as much as possible from changes in budgets and politics; states should employ the full set of policy levers at their disposal; and policies should directly support local reform efforts with proven effectiveness in addressing the experiences of students of different races attending the same schools.

Keywords: achievement gap; National Assessment of Educational Progress (NAEP); state education policies; hierarchical analyses.

## Acknowledgments

This research was supported by the U. S. Department of Education through a grant under the NAEP Secondary Analysis Program (Award Number R902B020004) and by the Educational Testing Service. A. Wang and F. Jenkins were employed by ETS during the course of the project. The authors thank A. Sedlacek (NCES), M. Nettles (ETS) and I. Lawrence (ETS) for their encouragement and support, as well as Stephen Raudenbush and Lauress Wise for their comments on an earlier version of the manuscript.

This report is the result of the efforts of many people and the authors are indebted to them. We wish to recognize the contribution of the late James Watts of the Southern Regional Education Board who helped us make contact with many of our experts. Of course we are especially grateful to the following individuals who responded to the questionnaire and answered questions from project staff.

**California**: Tej Pandey, Private consultant, Sacramento, California.

**Kentucky**: Scott Trimble, Kentucky Department of Education (retired), Skip Kifer, University of Kentucky, and Brian Gong, The National Center for the Improvement of Educational Assessment, Inc.

**Maryland**: William Schafer, University of Maryland (retired) and Paul Gammill, Baltimore City Public Schools (retired).

**Michigan**: Judy Lanier, Michigan State University (retired) , Sharif Shakrani, National Assessment Governing Board, and Scott Jenkins, Agate Software.

**New York**: Gordon Ambach, Council of Chief State School Officers (retired), John Bishop, Cornell University, and Benjamin Lindemann, New York State Department of Education (retired).

**North Carolina**: Gary Williamson, MetaMetrics, Inc., Richard Thompson, University of North Carolina, and Lou Fabrizio, North Carolina Department of Public Instruction.

**South Carolina**: JoAnne Anderson, South Carolina Education Oversight Committee and Sandra Lindsay, South Carolina Department of Education.

**Tennessee**: Douglas Wood, Columbia University, Karen Weeks, Tennessee State Board of Education, and Connie Smith, Tennessee Department of Education.

**Texas**: Chriss Cloudt, Texas Education Agency, and John Stevens, Texas Business and Education Council.

**Virginia**: Shelly Loving-Ryder, Charles Finley, and Pat Wright, all of the Virginia State Department of Education.

The following individuals reviewed and provided advice concerning the state profiles and data collection instruments: Douglas Tuthill, Private consultant, St. Petersburg, Florida; Margaret Goertz, Consortium for Policy Research in Education, University of Pennsylvania; Michael T. Nettles and Richard Coley, ETS; Patricia McAllister, Council of Graduate Schools; Carol Dwyer, ETS.

Technical support was provided by Terran Brown, Youn-hee Lim, Haiying Liu, and Terrell Strayhorn. Additional support was provided by Heather Rose Brooks and Juan Guzman. Superb editorial assistance was provided by Elizabeth Brophy. Any errors of fact or omission are those of the authors.

## Introduction

A critical and contentious issue in American education is the persistent gap in academic achievement between majority and minority students, especially Black students (Coleman et al, 1966; Jencks & Phillips, 1998; Kober, 2001). This gap has been documented at the national level at school entry (U. S. Department of Education [US DOE], 2002), at fourth and eighth grades (US DOE, 2004), and at twelfth grade (US DOE, 1995; US DOE, 1999).These reports make it clear that the gap at the means of the respective distributions is also reflected in the paucity of minority students at the highest levels of achievement.

Since the publication of the "Coleman Report,"[1] substantial resources have been devoted to "closing the gap" through such programs as Head Start, activities funded through Title I[2], as well as various programs sponsored by individual states. All too often, schools serving large numbers of disadvantaged students are underfunded and staffed with relatively new and/or underqualified teachers. Furthermore, economic disadvantage, high family mobility and rampant crime all can militate against success in school for these disadvantaged children. For recent reviews see Barton (2003) and Rothstein (2004). The federal and state governments have attempted to close the achievement gap through efforts aimed at raising learning standards and making state and local policies more coherent and comprehensive, thereby sending clearer messages to educators, students, and parents about what is expected. This is known as the standards-based movement. Furthermore, efforts to reduce funding inequities among school districts are intended to provide additional resources to districts with large proportions of students from poorer families, many of whom are minority. These efforts include Federal initiatives such as Title I, as well as initiatives undertaken by various states—often as the result of lawsuits (Odden & Picus, 2000).

[1] One of the most frequently cited studies about the influence of schools and students' home life on academic achievement was a 1966 study by James Coleman and others, *Equality of Educational Opportunity*.

[2] Title I (Part A) of the Elementary and Secondary School Act (ESEA) is the largest federal program in K–12 education, aimed at equalizing funding for high-poverty schools and funded at more than $11 billion in the 2003–04 school year.

Some reduction in the achievement gap since the 1970s in both reading and mathematics is indicated in the long-term trend component of the National Assessment of Educational Progress (NAEP) (Hedges & Nowell, 1998). However, results from main NAEP[3] from 1990 forward show that the gap has remained nearly constant. For example, in 8th grade mathematics at the national level, the gap between White students and Black students was 33 points in 1990 and 36 points in 2003 (US DOE, 2004). Moreover, the mean scale score in 2003 for Black students (252) falls at the 13th percentile of the distribution of scale scores for White students in 2003, which is an indication of the small overlap of the two distributions. At the same time, it is important to recognize that both groups improved their performance considerably over this period, White students by 18 points and Black students by 15 points. Although the Black-White achievement gap varies from state to state, the typical achievement gap at the state level is about the same magnitude as the achievement gap for the nation as a whole. Nonetheless, as we shall see below, some states have been more successful than others at closing the achievement gap, at least for some categories of schools.

As Federal and state authorities have implemented different strategies to raise achievement overall—and for specific groups—investigators have attempted to determine whether particular policies, or constellations of policies, are differentially successful in raising achievement and/or closing the gap. Evaluations of Title I (US DOE, 2001b; Kosters & Mast, 2003) or of the National Science Foundation (NSF) State Systemic Initiatives (SSIs) (Webb, Kane, Kaufman, & Yang, 2001) fall under this rubric. Some studies have focused on states' accountability efforts (Carnoy & Loeb, 2003), some on policies regarding teachers (Darling-Hammond, 2000), while others have considered a broader spectrum of reform initiatives (Swanson & Stevenson, 2002). Conclusions have been mixed and controversy is still the norm, particularly with respect to the long-term consequences of high-stakes testing (Amrein & Berliner, 2002; Carnoy & Loeb, 2003; Raymond & Hanushek, 2003; Braun, 2004; Desimone, Smith, Hayes & Frisvold, 2005; Nichols, Glass & Berliner, 2006).

This should not be surprising because such evaluations are a type of observational study of a small population of units (typically, states), with non-random allocation of "treatments." Consequently, causal inferences are not warranted. In addition, the actual causal pathways are complex, comprising many interacting systems whose dynamics can and do evolve through the years. Approaches that concentrate on a single policy, such as the introduction of high-stakes testing, or rely on a snapshot of the policy landscape at a particular point in time, cannot yield inferences that can be strongly defended.

Although most studies and public discussion rely on data reported at state or national levels, there is considerable evidence that there is more heterogeneity within states than among states (Raudenbush, Fotiu, Cheong, & Ziazi, 1996). This raises the possibility that patterns of achievement by race or trends in the achievement gap might present a different picture if they were viewed at lower levels of aggregation. (See also Bracey, 2003 on this point.) Analyses at levels below that of the state provide a better basis for examining the apparent effectiveness of policies targeted, for example, at raising the achievement of all students, including Black students; or at raising the achievement of certain types of schools, such as high poverty schools. On the other hand, studies of individual schools (or even a few schools) must confront the volatility and lack of generalizability attendant on small sample sizes.

The research presented here is an attempt to address some of these issues, by striking a compromise between state level and school level analyses. With constraints of cost and time in

[3] Main NAEP refers to that component of the NAEP program that is based on periodically revised curriculum frameworks. It is distinguished from the long term trend component of the NAEP program that is based on the original design established in 1969.

mind, we have chosen to examine the gap between Black students and their White peers, focusing on ten states in which Black students represent a substantial proportion of the public school population: California, Kentucky, Maryland, Michigan, North Carolina, New York, South Carolina, Tennessee, Texas, and Virginia.[4] The selection process is detailed below. Our analysis is distinguished by the examination of student achievement in subsets of schools (which we denote as "strata") that are determined by the average SES levels of their students.

We have two overarching research questions. First, has the Black-White achievement gap decreased from 1992 through 2000 in certain categories of schools in some of the ten states? And, in addition, do coherent and comprehensive state policies make a difference in closing the achievement gap?

It should be borne in mind that the size and trajectory of the Black-White achievement gap is completely determined by how the achievement of both Black and White students varies over time. Changes in the gap can result from different scenarios. For example, a reduction in the achievement gap may occur with scores in both groups increasing, but Black students experiencing a greater rate of improvement. Alternatively, White students' achievement may be essentially stagnant while Black students gain. Clearly, the interpretation and policy implications of a particular scenario should involve consideration of all the relevant trajectories. Accordingly, in this study, we present average results for each group separately, as well as for the difference in the averages (i.e., the achievement gap). In particular, we examine the possibility of an association between the reduction in the gap and the increase for Black students over the same period. We also examine the extent to which the variation in the gaps at different levels can be accounted for by characteristics of students and schools. Notwithstanding the complexities involved, we believe that the achievement gap is a critical indicator of the efficacy of our education system and merits attention in its own right. That the achievement gap appears to be generally resistant to policy interventions only serves to underscore the importance of examining its structure in greater depth.

Similarly, in order to address how effective state policies have been in closing the achievement gap from late 1980s through late 1990s, we must first understand how the policies have changed over time. Through our policy analysis framework, we identified five main policy levers: governance and the politics of reform, education finance, curriculum and standards, teacher quality, and assessment and accountability. For each of these levers, we identified some key reform-related questions for which we sought empirically based answers. To this end, we conducted, for each state, a comprehensive review of its policy history for the period 1988 to 1998. The review involved organizing and summarizing existing information and documentation relevant to the major policy dimensions, as well as extensive interviews with individuals knowledgeable about the state's initiatives in the education realm during this period. In particular, we relied on these experts to help us evaluate those aspects of states' policies that are more difficult to quantify but no less critical to their success: scope and quality, as well as coherence and consistency over time. Our approach is described in more detail in the Methods section.

With respect to achievement, we decided to examine student performance on State NAEP 8th grade mathematics, using results from the 1992, 1996, and 2000 administrations. This choice was made for a number of reasons. First, 8th grade mathematics represents the capstone of a state's testing program in mathematics mandated in the most recent reauthorization of ESEA, the No Child Left Behind (NCLB) Act. Second, there is a good deal of empirical evidence that math achievement, more so than reading, is influenced by teacher and school characteristics (Nye, Konstantopoulos, & Hedges, 2004; Dee & Keys, 2004). Finally, student proficiency at this level is

[4] In what follows, we will often refer to the states by their usual two letter abbreviations.

predictive of subsequent success in high school and participation in post-secondary education. (See the report of the US DOE [2001b] and the references therein.)

Using auxiliary information from the NAEP administrations, we categorized schools in the NAEP sample as higher poverty or lower poverty, based on the percentage of students in the school who were eligible for free or reduced price lunch. Schools with 50% or more of eligible students were classified as higher poverty. The others were classified as lower poverty. In what follows, we refer to these two categories as strata. For the period 1992 to 2000, we computed the changes in 8th grade NAEP mathematics scores for White students, Black students and the differences between them (the Black-White achievement gap) at the state level and at the stratum level within the state.

To facilitate the interpretation of the results, we first set the magnitudes of the Black-White gaps within strata against the general background of between state differences (the usual focus of both policy-makers and the public) as well as against the typical sizes of between stratum differences within states. We then compared states in terms of the changes over the period in the Black-White achievement gaps within each stratum. (This is a reasonable strategy since the definition of the stratum is the same for all states and across administrations.) Contrasts between the higher and lower poverty strata within states were also examined.

Next, we effectively restricted our attention to schools whose NAEP samples included both Black students and White students. For those schools, we employed hierarchical linear models (HLM) to partition overall achievement variance into between-student, within-school and between-school components. We then augmented the basic variance components model with student and school characteristics. Thus, for each stratum within a state, we were able to estimate the size of the Black-White achievement gap for those students attending the same schools—usually denoted as a "pooled within school" estimate. Moreover, we were able to determine how much of that gap could be accounted for by other student characteristics, such as student socioeconomic status (SES). Although of secondary interest, we also estimated how much of the between school variance could be accounted for by school characteristics, such as their demographic make-up.

We developed a summary description of the achievement record of each state and categorized states on the basis of those summaries. We considered both absolute gains by Black students as well as progress in closing the achievement gap between Black and White students. Separately, we categorized states with respect to their policy histories, first considering each policy lever separately and then a derived overall policy score. Finally, we juxtaposed the two achievement categorizations against the policy categorizations, identifying patterns of interest.

Of course, we are mindful of the fact that, notwithstanding the extensive amount of data analyzed, the observed patterns can only lead to tentative conclusions regarding the effectiveness of particular policies or strategies. The arguments must necessarily be indirect and circumstantial. Nonetheless, we believe that our approach has yielded insights that can be helpful to states as they move forward with their own initiatives and wrestle with the requirements of NCLB. In any event, the results we have obtained can serve as a baseline against which to compare the success of these states over the next decade in improving achievement and closing the Black-White gaps.

The article is organized as follows: In the next section we provide an extended review of the literature. We then describe our methods, followed by our policy analysis and then separate sections containing the basic descriptive results and the multi-level analyses. The next section presents the linking of the policy and quantitative analyses. The paper concludes with a short discussion of findings and implications.

# Literature Review

## *Concerns with U.S. Public Education*

There have always been criticisms of the public education system and, correspondingly, a plethora of attempts to remedy the purported problems. This dynamic has been well documented by Tyack and Cuban (1995) and, more recently, by Ravitch (2000). Two key issues are appropriate goals for the system and effectiveness in achieving those goals. Particularly in the last decade, some critics have questioned whether, even with the provision of additional resources, public education is flexible enough to evolve over time to meet increased demands, including better serving all students. Advocates of charter schools and vouchers maintain that without competition and real alternatives, efforts to implement wide-ranging improvements in public education are doomed to failure (Hoxby, 2001).

Over the last twenty years or so, there has been growing concern about how U.S. students—and adults—perform in comparison to their peers in other nations. In a recent analysis of data from the International Adult Literacy Survey, Sum, Kirsch, and Taggart (2002) characterize U.S. performance by the phrase "mediocrity and inequality." They point out that, on a number of dimensions, the U.S. places at or below the median of 17 developed nations and also exhibits much greater variability than any of the other nations. Although Sum et al. are concerned with the performance of adults, the same characterization applies to the achievement of in-school populations.

Comparisons from international assessments, such as the Third International Mathematics and Science Study (TIMSS) and Third International Mathematics and Science Study-Repeat (TIMSS-R) (US DOE, 2000) make it clear that overall U.S. performance is far from "world class." Results of the first assessment carried out under the auspices of the Program for International Student Assessment (PISA) were released by U.S. Department of Education (2002). Among the 28 Organization for Economic Co-Operation and Development (OECD) nations that took part in this assessment of 15-year-olds, the U.S. scored about average on reading, math and science. Bracey (2002) points out that if U.S. students were disaggregated by race/ethnicity, White students would place 7th, while both Black and Hispanic students would place 27th. Obviously the wide disparities among demographic subgroups contribute to the "excess variability" noted above.

The magnitude and persistence of these disparities is rightly regarded as problematic both for our long run economic competitiveness and the health of our democracy (Friedman, 2005). They are particularly troubling given the projections of the future demographic composition of the U.S. population. For a broader and more optimistic view, the U.S. Department of Education report (2001a) on educational achievement and Black-White inequality is instructive. Among the findings are that Black-White disparities in college attendance, employment, and earnings are reduced, eliminated or even reversed if individuals are matched on prior educational achievement. However, as long as there are substantial differences in the distributions of achievement, overall differences between Blacks and Whites will remain large and problematic.

Bracey's (2002) observation regarding disaggregated results is but one example of the problems in reporting data at one (usually the highest) level of aggregation. Raudenbush, Fotiu, Cheong, and Ziazi (1996) make a similar point with respect to comparisons among states based on NAEP data. The latter study also demonstrates that there are substantial differences among states, and among race/ethnic groups within states, in home environment and in learning opportunities in school. Not surprisingly, these differences are strongly associated with differences in achievement.

In view of the American tradition of local control and, in many states, heavy reliance on the local property tax base to fund schools, there are serious questions about whether mechanisms can be found to provide adequate funding to impoverished districts with poor records of achievement. In many states, litigation has resulted in court mandated reforms—but these have often been stymied or poorly implemented, with the result that little progress in equalizing per pupil expenditures has been made.

## *The Achievement Gap*

The persistent differences between Blacks and Whites in both educational achievement and educational attainment have been well documented in a number of sources. A brief historical introduction to the problem, as well as a thoughtful review of relevant issues, is provided by Jencks and Phillips (1998). Hedges and Nowell (1998) analyze data from six surveys and present evidence that the achievement gap has narrowed somewhat since 1970 but that the differences remain substantial. Moreover, they note that "Blacks are hugely underrepresented in the upper tails of the achievement distribution, and this representation [gap] does not seem to be decreasing" (p.167). [5]

These conclusions are echoed in a recent report by Barton (2002) for the National Education Goals Panel in which he analyzes data from both National and State NAEP over the period 1990 to 2000. He reports, for example, that over that decade, the gap in 8th grade mathematics went up in 15 states and down in 14 states; however, only two of the 15 were statistically significant and none of the 14 were. The U.S. Department of Education (2001) report already cited is a comprehensive study of the differences between Blacks and Whites in educational achievement as well as in a variety of educational and economic outcomes. One of the conclusions drawn is that, "The black-white mathematics gap differed in size across grades, in a manner consistent with, but not necessarily demonstrating, a narrowing of the gap during elementary school, followed by a widening of the gap during junior high school and little change during senior high school" (p. v). It should be noted, however, that these findings are based on stitching together results from a number of cohorts spanning different grade ranges. Thus, they could be affected by unmeasured between-cohort differences. (See Ludwig, 2003, for further comments.)

There is a considerable literature on the non-school correlates of achievement and myriad partial explanations for the existence of the achievement gap. Miller (1995) presents an exhaustive analysis of the correlations between social class and socio-economic status on the one hand and academic achievement on the other. The correlations are strong for all race/ethnic groups and for all standardized measures of achievement. Moreover, he documents the wide variations among race/ethnic groups in education-relevant resources and their association with differences in achievement. He explains Coleman's (1966) conclusion that family variables, rather than school variables, account for more of the achievement differences among race/ethnic groups as follows: "The variations in home-based, formal-schooling-derived resources that have been intergenerationally accumulated by families are greater than the variations in education-relevant resources that society is investing in the current generation of children directly through the schools" (p. 119).

[5] Of course, for typical test score distributions, a substantial gap at the means of the individual distributions is accompanied by highly disproportionate representation in the tails of the combined distribution.

Miller also addresses the troubling finding that even controlling for social class (using available measures such as parental educational attainment), there are considerable differences among race/ethnic groups in academic achievement in kindergarten through 12th grades. However, these differences parallel the findings "that there are currently very consequential differences in the amounts of human capital possessed by young white adults and their African American and Latino counterparts and that these variations exist at most educational attainment levels" (p. 170). Again, these disparities appear to be the consequence of differential rates of accumulation of human capital over many generations. From a policy perspective, the inescapable conclusion is that the closing of the achievement gap will only happen over generations. The caution for the methodologist is that available measures of family resources underestimate group differences. This bias must be taken into account when interpreting the results of any analysis.

Phillips, Brooks-Gunn, Duncan, Klebanov, and Crane (1998) offer an accessible account of the issues as well as a sophisticated analysis of data from the Children of the National Longitudinal Survey of Youth. They investigate a number of factors related to family background and parenting styles and habits and conclude that perhaps as much as two-thirds of the test score gap can be accounted for by these factors. They acknowledge, however, that other authors (Herrnstein & Murray,1994; Hedges et al., 1998) reach a figure closer to one-third. The latter estimates employ a more restrictive definition of socio-economic characteristics. Note that Phillips et al. (1998) augment the usual family variables with a number of other educationally relevant variables that reduce the misspecification described by Miller (1995). They are also at pains to point out the potential confounding of genotype and home environment, which complicates both interpretation and prediction.

More recently, Barton (2003) continues this line of analysis. Based on a review of the literature, Barton identifies 14 correlates of academic achievement that involve factors related to health, housing, nutrition and school quality. He then provides data to estimate the gap between minority and majority students. For each factor, the gap favors majority students and, in most cases, the gap is very substantial. A plausible conclusion is that the achievement gap is but one consequence of these differences among students and, moreover, that school reform alone is unlikely to eliminate the gap. For a slightly different view, see Lee (2002).

Indeed, school effects have been much studied since Coleman et al. (1966), with the more recent studies examining the contributions that schools may make to the achievement gap. For example, Phillips, Crouse, and Ralph (1998) conduct a meta-analysis and report that, "Black students who start elementary school with the same test scores as the average White student learn less than the average White student between the first and twelfth grades" (p. 257). They further note that "…our results imply that neither differences between the schools that Blacks and Whites attend nor differences in their socioeconomic status suffice to explain why Blacks learn less than Whites with similar initial skills" (p.267). Ferguson (1998) investigates the differential effects of grouping and tracking and concludes that they cannot account for much of the achievement gap. Clearly, no single factor can account for the large and persistent test score gaps that have been observed and the complexity of the problem may be the principal reason that it has proven so refractory to amelioration, at least on a large scale.

## *Education Reform*

In addition to the contributions of families and communities, schools can make a difference in closing the achievement gap. Throughout the 1990s, there was a shared belief that low student achievement in public schools was primarily the result of low standards, incoherent and fragmented policies, and poor use of resources (Corcoran, 1997; US DOE, 2001c). As a result, the 1994 reauthorization of the Elementary and Secondary Education Act (ESEA), the Goals 2000 program, and the National Science Foundation's State Systemic Initiatives all focused on top-down reform. States were asked to set higher standards and expectations for all students; create more coherent and aligned assessments and curricular frameworks; implement an accountability system with rewards and sanctions; and change teacher certification requirements to enable teachers to prepare students to meet the new standards (Corcoran, 1997).

States are constitutionally responsible for elementary and secondary education and generally play a major role regulating education, although there is considerable variation across states in the balance between centralized and local (district) control. Where there is sufficient political will, however, every state has the power to promulgate new rules, establish standards, mandate accountability measures, and impose sanctions or rewards. Nonetheless, state-initiated reform efforts are shaped in part by traditions of local control and by the challenge of effectively driving change down through myriad local bureaucracies. Federally-initiated reforms have to contend with constitutional issues, although the most recent re-authorization of the ESEA, the No Child Left Behind Act, appears to have had a dramatic and immediate impact on states' policies through the threat of withholding federal funding in the event of non-compliance.

During the 1990's, many state reforms focused on disadvantaged students and some directly on the achievement gap—with modest, if any, success (Brady, 2003). Clearly, there are many possible explanations to account for this discouraging record, including the challenge of maintaining coherence, consistency, and adequate resources in a dynamic political environment, as well as the general difficulty of propagating "top-down" reform without substantial attenuation at the classroom level. Another view is that schools have multiple missions that conflict with one another (Christensen & Karp, 2003). Toward the end of the decade, a number of reports examined the achievement gap and offered recommendations on actions that policy-makers, school officials, parents, and others could take to improve achievement overall and to reduce the gap (College Board, 1999; Center on Education Policy, 2001).

## *Evaluating the Effectiveness of Reform Efforts*

The initiatives taken at the federal and state levels, such as Title I and the NSF's State Systemic Initiatives (SSIs), are based on assumptions that schools can make a difference in student learning, notwithstanding the substantial influence of family and community characteristics. (Of course, there is a second, implicit, assumption that governmental policies can constructively influence the practice of schools.) There is certainly considerable anecdotal evidence of schools that have made a difference (Education Trust, 1999, 2001; Cawelti, 1999), but large-scale analyses have yielded inconsistent results with continuing controversy ever since the publication of the Coleman Report in 1966.

As part of the 1994 reauthorization of the ESEA, Congress mandated the national evaluation of Title I to examine the progress of students whom the program was intended to benefit, as well as the implementation of key provisions of the program. The evaluation found that Title I reached more than 12.5 million students, many of whom were from the highest-poverty schools. Although Title I specifically targeted students in poverty, the impact of Title I program on student achievement could not be easily disentangled from the contributions of other factors, including the state and local reform efforts which Title I was designed to support. Moreover, results from state assessments and from NAEP present somewhat contrasting pictures of Title I's success in narrowing the achievement gap. State assessments generally indicate some progress in narrowing the achievement gap while the results from the long-term trend component of NAEP indicates a slight widening of the achievement gap from the late 1980s to 1999 (U.S. DOE, 1999, 2001c)[6] .

The vision outlined by NSF's SSI was even more ambitious. It included high standards, along with aligned curriculum, pedagogy, and assessment. Proponents of systemic reform believed that high standards for student learning should form the basis for the alignment of all policies, practices, and resources throughout the educational system. They posited that improvement in student achievement requires coherent policies and coordinated resources that are designed to communicate a clear vision of what students should know and be able to do (Zucker, Shields, Adelman, Corcoran, & Goertz, 1998; Corcoran, 1997, Clune, 1998).

A total of 24 states and Puerto Rico participated in the SSI program in the 1990s. These states experimented with a variety of strategies. Evaluations of SSIs have found that half of the SSI states showed some impact on classroom practice. Moreover, achievement gains were higher in states that had intense professional development focusing on curriculum and materials (Blank, 2000, Zucker et al. 1998). However, researchers concluded that assessing the extent to which an SSI achieved the goal of closing the achievement gap was challenging, as the capacity to do so varied greatly from one SSI to another and, furthermore, most states were not able to ensure equal implementation of SSI activities across all schools. In fact, only a few SSIs attempted to change or restructure the professional development system itself to ensure that all teachers were given access to high quality training. However, it does seem that states that implemented a focused SSI or standards-based reform were able to make a difference in their teachers' self-reported classroom practice.

It should be evident that the evaluation of such broad reform efforts must confront a number of challenges. These include difficulty in establishing clear specifications and measurement rubrics for different policy components, the characterization and tracking of changes in these components over time, and the ability to disentangle the effects of the intended policies on targeted outcomes from other factors. Most researchers agree that no single study or approach can provide a definitive answer to any realistic policy question. The issues involved are too complex and the limitations of a particular approach too numerous to fully exclude plausible competing explanations. Instead, it is necessary to triangulate among cross-sectional, longitudinal, and experimental studies. Unfortunately, the time frame for such a comprehensive strategy usually does not conform to the needs of decision-makers.

Although analytic methodologies have increased in sophistication and the databases have become richer and more comprehensive, there is still a frustrating lack of consensus on "what works." The key is to try to exploit the links between variation in practices and differences in outcomes. Presumably, once unusually effective (or ineffective) schools and teachers have been identified, further investigation is required to determine the specific practices that may be responsible. Such intensive study is time-consuming and expensive (see Klein et al. [2000] and the references therein).

Turning to the evaluation of the effectiveness of state actions, there are many policy differences among states that can be considered. Of course, much depends on the outcome measure selected. Barton and Coley (1998), for example, follow a cohort of 4th grade students in 1992 who attended the 8th grade in 1996 and estimate the gains for each state, based on State NAEP results. As Camilli (2000) points out, Texas ranks below the median on this measure, rather different from its

[6] Discrepancies in trends between a state assessment and NAEP can be due to many factors including differences in the assessment frameworks, test content, administration protocols and student motivation. For further discussion, consult Thissen (2005) and Koretz (2005).

rank as number one in other accounts of the "Texas Miracle" (Haney, 2000). A recent update can be found in Coley (2003). It is not surprising to find differences in results between cross-cohort and within-cohort analyses (Braun, 2004).

Certainly, participation in the State NAEP assessments offers states a number of evaluation alternatives. While some of the concerns mentioned above with respect to state tests are allayed with State NAEP, others remain and new ones appear. As an example of the former, there are worries that some states may try, at least to some extent, to exclude many low-performing but eligible students in order to improve states' reported performance. On the other hand, because participation in NAEP is not mandatory, the representativeness of both the school and student samples must be closely monitored. Furthermore, inasmuch as NAEP does not report scores at the student level and is not considered a high stakes test, variation in student motivation (within and between states) is a potential source of bias that is difficult to quantify.

Nonetheless, State NAEP represents a rich database for analyses at the state level and below. By now there have been a sufficient number of administrations to justify the exploration of the possible effects of reform efforts. There appear to be two main strategies in the use of State NAEP. One relies on aggregation of data to the state level and, depending on the nature of the models adopted, proceeds to draw inferences about the state or about comparisons among the states. The other employs so-called multi-level or hierarchical linear models (HLM) (Raudenbush & Willms, 1995; Raudenbush, 1988; Raudenbush & Bryk, 2002) to develop more detailed pictures of the relationship between achievement and other measured variables as a basis for inference.

A report by the Education Intelligence Agency (Antonucci, 1999) offers a relatively simple but interesting example of the first approach. He first computes a cumulative score for each state by combining its scaled scores on Grade 4 math and reading as well as Grade 8 math and reading obtained from the 1996 and 1998 State NAEP Assessments. He then examines, for example, how state rankings change when attention focuses on the performance of the states' Title I students. Texas, which is near the median in the overall ranking climbs to the top in the latter one, while California remains mired near the bottom on both. He also compares states on the differences in scores between students whose teachers had a masters degree and those whose teachers had a bachelors degree, In each state, the comparison favors the former but in only a third of the states was the improvement per test of practical interest. In this case, California and Texas were adjacent, and at the median. While the specter of confounding is ever-present, Antonucci does make some interesting observations that lead to hypotheses to be further investigated.

More recently, Lee (2002) has considered trends in Black-White and Hispanic-White score gaps in reading and math using long-term trend NAEP. He observes that the "…achievement gaps narrowed in the 1970s and 1980s but then stabilized or widened in the 1990s" (p. 3). He then attempts to correlate these patterns with long term trends in various measures of socioeconomic, educational and cultural conditions. He argues that the latter trends cannot account for the widening of the gaps observed during the 1990s and that it will be "… necessary to investigate simultaneous changes across a broad range of factors from multiple data sources and to examine their interactive, joint influences on the achievement gap" (p.10).

Darling-Hammond (2000) considers state averages on reading and math (1990, 1992, 1994, 1996) as well as state gains in Grade 4 math (1990 to 1996 and 1992 to 1996) and state gains in Grade 4 reading (1992 to 1994). In addition to State NAEP, she draws on the 1993–94 Schools and Staffing Survey and data from a 50-state survey of teacher policies. The goal of the investigation is to link state achievement to state initiatives targeting the improvement of teacher quality. A variety of analytic techniques are employed, including comparing similarly situated states with different gain records and adjusting state results for differences in contextual variables such as student poverty. This effort is noteworthy for combining quantitative analysis with a comprehensive set of case studies. While recognizing both that states often engage in multiple reform initiatives and that aggregate estimates necessarily miss important within-state variation, she concludes that, "…states interested in improving student achievement may be well-advised to attend, at least in part, to the preparation and qualifications of the teachers they hire and retain in the profession" (p. 35).

Grissmer, Flanagan, Kawata and Williamson (2000) adopt an approach that is analytically more complex. They employ data from the 1990 U.S. Census and from NELS to augment the family data collected by State NAEP, with the rationale that NAEP does not collect some important variables and that some of the student self-report data NAEP does collect is of low reliability. Their starting point is the incomplete matrix of 44 states crossed with the seven 4th grade or 8th grade math and reading tests that were administered between 1990 and 1996 (inclusive of the end years). A cell entry is the (normalized) score of a particular state on a particular test. There were 271 cell entries in all.

Grissmer et al. (2000) proceed to develop and fit a number of different models in order to estimate such quantities as annualized state gains in reading and math and, "to determine whether trends and differences in scores across states for students from similar family backgrounds can be statistically linked to differences in state educational system characteristics that are resource intensive…" (p. xx). In a separate set of analyses they also attempt to estimate the cost-effectiveness of different reform policies. Among other things, they conclude that 1) most states experienced statistically significant gains in 8th grade math over the period 1990–96; and, 2) policies that emphasize enrolling more children in public pre-kindergarten programs, reducing pupil-teacher ratios in the lower grades, and providing teachers with adequate resources appear to be effective in improving achievement—at least in states with large numbers of disadvantaged students and whose educational expenditures were below the national average at the beginning of the period.

More recently, Desimone et al. (2005) developed a five-dimensional framework for describing state education policies, as well as protocols for developing descriptors of each state with respect to the framework. They then used multiple regression to relate changes in state performance on NAEP 8th grade mathematics, from 2000 to 2003, to various combinations of state policy descriptors. They studied the NAEP composite score, as well as subscores for procedural knowledge, problem solving and conceptual understanding. This approach was most successful in accounting for variation in states' procedural knowledge averages in 2003 in terms of four policy descriptors, states' procedural knowledge averages in 2000, and the interactions among these five predictors (adjusted $R^2 = .5$). There was some evidence that gains were greater for states with lower average scores in 2000.

Nichols et al. (2006) contribute to the long-running debate about the efficacy of high-stakes testing in improving student learning, as measured by test scores. Their contribution is two-fold. First they develop and validate a new indicator, the Accountability Pressure Rating (APR), which ranks states with respect to the pressure exerted on school systems through the implementation of high-stakes testing and its associated consequences. They also introduce a second indicator (EPR) that tracks meaningful changes in the APR over time. Second, they carry out a series of analyses that examine the pattern of relationships between the APR or the EPR and a variety of criteria, including NAEP gains in math and reading over different time periods. Analyses are carried out both for the entire student population and for specific subgroups. The scope of the study and the heterogeneity of the results make it impossible to provide a simple summary. Suffice it to say that there is some evidence that accountability pressure is related to student gains in grade 4 mathematics, but neither for grade 8 mathematics nor for reading in either grades 4 or 8.

Strategies that employ HLM typically focus on the analysis of test scores within a state and take account of the fact that NAEP student samples are clustered by school. They provide a more detailed picture of the structure of achievement in the state as well as a more defensible partitioning of error variance. An excellent example of the application of HLM to important policy issues is provided by Raudenbush et al. (1996). They employ data from the 1992 State NAEP Assessment in 8th grade math to study state-to-state variations in achievement as well as differences between majority and minority students. They find that a combination of home environment indicators and school learning opportunity indicators account for most of the considerable between-state variation in 8th grade math achievement. Interestingly, they also examine differences among states in the overall availability of learning opportunities, as well as social and ethnic inequalities in access to those opportunities. Not surprisingly, here, too, they find substantial variation among states. The story that emerges from this analysis offers much more to policy-makers and educators than do simple state rankings by overall achievement.

Swanson and Stevenson (2002) used hierarchical linear models to evaluate the impact of the national standards-based reform on classroom instructional practices, using data from the 1992 and 1996 state National Assessment of Educational Progress (NAEP) in eighth grade mathematics. They found that the states with standards-based reform were more likely to have teachers' self-reported classroom practices that were consistent with the standards-based models of mathematics education.

With cross-sectional observational studies like NAEP, inferences about policy impact must be made indirectly and circumspectly. Grissmer et al. (2000), for example, conclude that certain residual patterns in the data not accounted for by the model can be plausibly linked to particular reform initiatives. They acknowledge that other explanations are possible and could be developed if more comprehensive data were available. It is worth noting that the conclusions of Grissmer et al. (2000) concerning the impact of state policies do not fully accord with those of Darling-Hammond (2000) with respect to teachers nor with those of Hanushek and his collaborators (Hanushek, Rivkin & Taylor, 1996) with respect to the impact of resource investments.

In view of the current level of disagreement, these authors as well as Raudenbush et al. (1996) concur that further study is warranted. Even for a particular type of survey data, like State NAEP, there are tradeoffs among different kinds of models. As indicated above, HLM yields a more detailed picture of the topography of achievement and its correlates, but at the cost of having to grapple with sampling variability and the quality and extent of the relevant contextual data. Indeed, Grissmer et al. (2000) make a cogent argument that an important advantage of state-level analyses is that it is possible to augment NAEP data with high quality data from other sources that should reduce model misspecification. They also include an extensive discussion of the problem of aggregation bias. We have taken these issues to heart in the development of our research design.

## *Framework for Evaluating State Education Policy*

In our view, a rigorous investigation of the link between state education policy and student outcomes requires that policy descriptions be constructed around a framework. If successful, such a framework helps to organize data collection and analysis. Equally important, it facilitates meaningful comparisons among states. There have been a number of attempts to develop useful policy frameworks and they are described in brief below.

In their seminal paper that helped to launch the standards-based reform movement, Smith and O'Day (1991) proposed a dual strategy to influence student achievement: state leaders would initiate a coherent and directed reform centered around an instructional guidance system for schools and districts, while local decision makers would have flexibility in adopting and adapting system components to maximize the quality of classroom instruction. The components of an instructional guidance system include curriculum standards and frameworks, curricular materials, assessments, professional development, teacher licensing, teacher re-certification, technical support, accountability, and organizational capacity building (Smith & O'Day, 1991).

This theory of action has been elaborated by many others since. The central argument is that if governments set standards for student performance, and adopt aligned policies for curriculum, assessment, accountability, and governance, then educators will alter their practices, and student performance will improve as a consequence (Corcoran, 1997). The main strategy of systemic reform intended to close the achievement gap is to raise expectations and standards for all students, to align assessments and curriculum with the standards, and to rationalize the evaluation of teaching and learning.

One version of a standards-based policy framework was posited by Swanson and Stevenson (2002) in their quantitative evaluation of the impact of the national standards-based reform on classroom instructional practices. Specifically, they were interested in the correspondence between state-level activism in content standards, performance standards, and aligned assessments and professional standards, on classroom instructional practices. They developed a composite measure based on a set of 16 teacher-reported classroom instructional activities as the dependent variable, which were collected from a series of studies conducted by the Council of Chief State School Officers (CCSSO) on state policy actions.

Clune (1998), in the context of evaluating nine NSF statewide systemic initiatives, proposed a framework for a state's standards-based reform efforts. The framework has four related aspects (i.e., standards-based reform, standards-based policy, standards-based curriculum, and standards-based achievement), each with a number of secondary elements. He argued that each aspect should be rated with respect to breadth and depth, where breadth referred to the number of elements involved and depth referred both to the quality of the implementation and to the strength of influence. This framework was used to evaluate the effectiveness of nine statewide systemic initiatives based on rich case studies.

Desimone et al. (2005), cited above, review some of the relevant literature and propose a policy framework comprising five attributes: consistency, specificity, authority, power and stability. These are described and operational definitions provided. Of particular interest, they examine empirically the relationships among the attributes for the states in their sample. While the article appeared after the present work was completed, there is certainly some overlap in our frameworks. In particular, we also argue that consistency and stability are important aspects of state policy and should be incorporated in any study that seeks to link policy and education outcomes.

Edwards and the staff of *Education Week* (1997, 1998, 1999, 2000) developed an alternative framework for grading state efforts in raising student achievement in their Quality Counts reports. This framework addressed five aspects of state policies: achievement; standards and assessment; quality of teaching; school climate; and resources. They believed a successful public school system should have these characteristics: high standards in English, math, science and history for all children and assessments that align with those standards; teachers whose primary focus is on student learning and who possess the knowledge, skills, and commitment to teach to higher standards; schools that are organized and operated in a way that encourages and supports teaching and learning; adequate funding distributed equitably to all children and focused on the functions that matter; and all students achieving at high levels and engaged in challenging intellectual work. They drew their state-level data from such organizations as the U.S. Department of Education, the Education Commission of the States, the American Federation of Teachers, and the Center for Education Reform.

## Methods

### *Preliminaries*

*Selection Criteria for Participating States*

We adopted three criteria for selecting states for the study. First, the chosen states must have participated in the 8th grade State NAEP mathematics assessment in the years 1992, 1996, and 2000. Eighth grade math was chosen because it represents the capstone of the state testing programs in mathematics mandated in the NCLB Act. There is also a significant amount of empirical evidence that math achievement, more than reading, is influenced by factors that are under state policy control, such as teacher and school characteristics. Moreover, a student's proficiency in math at the 8th grade level is predictive of subsequent success in high school and participation in post-secondary education. Second, the selected states' public school population had to have encompassed a minimum of 10 percent Black students. While we recognize that all disadvantaged groups confront barriers to achievement, we focused on Black students so that we could concentrate our attention on a manageable number of states. Finally, at least some of the states selected had to be considered "bellwether states"[7] on educational policy over the period of interest.

Four states appeared to meet these criteria: Kentucky, North Carolina, Tennessee, and Texas. Initially, California was not included because its public school population did not encompass a minimum of 10 percent Black students. We decided to add it to the study both because it has a large number of Black students and because it is prominent in education reform. We augmented our state sample by including Maryland, Michigan, New York, South Carolina, and Virginia. These states all had three administrations of State NAEP mathematics assessment and had a minimum of 10 percent black students in their public school populations. We believed these states could provide useful contrasts to the other states.

*Analysis of State Policy Data*

We used the case study approach to document state educational policies over the period 1988 through 1998. Our research questions were: What policies were in effect during this ten-year period? What was the sequence of these policies? In the aggregate, how coherent and consistent were these policies?

In order to address these questions, we developed a framework for policy analysis that guided the development of our state profile, a questionnaire on state education policy, and a follow-up interview protocol. The framework builds on the earlier work just cited and comprises five components, which we refer to below as "policy levers." The state profile and questionnaire were revised several times, based on comments from internal and external reviewers. We submitted the final versions of the profile and questionnaire to at least two experts in each state,[8] asking them to review the profile, respond to the questionnaire and to participate in an interview. We then triangulated multiple sources of evidence to create an extensive state summary that was eventually reduced to a one-page synthesis for each state.

[7] We define bellwether states as those states that are generally considered to have been leaders in systemic education reform.

[8] Only one expert from California was interviewed.

After evaluating the state syntheses against our framework, we classified the states into one of three ranked categories, based on our judgment as to the extent to which state policies over the period would lead us to expect substantial improvement in student test scores from 1992 to 2000. The categorization process was carried out for each of the five policy levers, as well as for an overall policy score. The next section describes each of these steps in more detail.

### *Establishing the Framework for Policy Analysis*

The policy analysis framework was developed through an iterative process. We began by reviewing both the theory underlying standards-based reform and the efforts to evaluate the theory. We drew from the work of Smith and O'Day (1991), Swanson and Stevenson (2002), Clune (1998), and the Quality Counts framework (Edwards, 1997, 1998, 1999, 2000). Although each framework had much to recommend it, one concern we had was that none appeared to cover the full range of policy levers available to states. Another concern was that they did not address either how well policies were carried out over time or how strongly aligned they were. Our premise was that for any educational reform to have an impact, it must be developed, implemented, and refined through time (Brady, 2003).

Our framework comprises five policy levers through which states can regulate and monitor their education system: governance, education finance, curriculum and standards, teacher quality, and assessment and accountability. We further posit that an evaluation of state reforms must take into account the quality of implementation, as well as their coherence and consistency over time.

In other words, our framework portrays state policy as the means by which a state can effectively implement its statutory and regulatory authority (governance) through its ability to raise, distribute, and spend money on public education (education finance). These financial inputs influence how much instructional and other resources are available to students and their teachers. Another way for states to influence educational policy and academic achievement is through the specification of what students should learn and be able to achieve (curriculum, standards and accountability), and who should teach them (teacher qualifications and quality). Finally, the degree to which states have coherently and consistently developed and implemented desired policies is instrumental to their success.

### *Developing the State Profile, Questionnaire and Interview Protocol*

Employing the framework described above, we developed a sample profile of one state using information obtained from over thirty sources including the Council of Chief State School Officers (CCSSO), the National Center for Education Statistics (NCES), *Education Week Quality Counts*, the Education Commission of States, and the Consortium for Policy Research in Education. We also consulted the published research literature as well as a large number of project reports. We then shared this sample profile with five advisors.[9] Based on their comments, we revised the profile and shared the second draft with them. We then developed the final version of the state profile (Appendix A), which was subsequently used for all ten states.

Using a similar iterative process, we developed a questionnaire on state education policy (Appendix B) and a follow-up interview protocol (Appendix C). We structured the questionnaire

[9] These were Margaret Goertz, Co-Director of the Consortium for Policy Research in Education; Douglas Tuthill, private consultant, St. Petersburg, FL; Michael Nettles of Educational Testing Service; Rich Coley of Educational Testing Service, and Patty McAllister of the Council of Graduate Schools.

and the interview protocol around the five policy levers. Through these instruments we were particularly interested in obtaining information with respect to issues relating to coherence, consistency, the quality of implementation, perceived positive and negative effects, and lessons learned.

### *Selecting State Experts*

Except for California, for which we only conducted one interview, at least two individuals in each state were identified on the basis of their knowledge of state policy actions over the period 1988 to 1998, and their willingness to participate in this study. Typically, they work or worked in the state's department of education or related agencies, regional education laboratories, or policy research institutions. They were sent the draft profile and the questionnaire and asked to review and correct the profile, fill out the questionnaire, send the materials to the researchers, and participate in a 45-minute phone interview. The level of cooperation was excellent. The experts substantially contributed to our understanding of the coherence, consistency, and quality of state policy actions over the period from 1988 to 1998. They also provided references to relevant materials and suggested other respondents.

### *Developing State Summaries and Ranking the States*

Once all the state information was in hand, we triangulated the information and compiled a comprehensive "state summary" for each state. (An example for Kentucky is contained in Appendix D.[10]) By focusing on those issues that best differentiated the states, we further reduced the summary to produce a one-page synthesis for each state. We then evaluated each state synthesis in comparison to the others and then classified the states into one of three categories based on our judgment as to whether state policies over the period would lead us to expect substantial improvement in NAEP results from 1992 to 2000.

## *Stratification of Schools*

An important goal of this study was to develop parsimonious descriptions of the patterns of achievement of Black students and White students, paying special attention to trends in the achievement gaps. We were particularly interested in describing those trajectories at a level below that of the state as a whole. Of course, we recognized that we would have to balance the construction of more homogeneous groups of schools from a state's NAEP sample against the increased variability due to smaller sample sizes. This was especially problematic as the number of schools in a state sample could fall well below the target of 100 to 105 schools.

The data for all analyses were taken from NAEP Restricted Use Data Products[11] and were organized into 10 x 3 = 30 data sets, one for each combination of state and administration. In each data set, the information for an individual student consists of five plausible values on the NAEP 8th grade mathematics scale, an extensive set of student/teacher/school background variables collected during the NAEP administration, and the variables employed by Westat in developing the sampling

[10] The full set of state summaries is available from the first author.

[11] NAEP Restricted Use Data Products are available to qualified researchers by license from the National Center for Educational Statistics, U. S. Department of Education.

plan for the state. For further details, see Chapter 3 of Allen, Jenkins, Kulick, & Zelenak (1997). Subsequently, we linked the NAEP files to the Common Core of Data (CCD) files of the same year using NCES school codes. [12]

Inasmuch as our interest was in studying trends in achievement for groups of schools within a state, we began by categorizing schools with respect to three characteristics: Type of location (TOL), percentage minority (%Min) and percentage of students eligible for free or reduced price lunch (the lunch-program percentage). The first two characteristics are obtained directly from the Westat files and consist of up to eight and up to four categories, respectively.

For the lunch-program percentage, it was decided to develop two categories: Schools with less than 50% of students eligible and schools with 50% or more eligible. These categories are denoted as Stratum 1 (S1) and Stratum 2 (S2), respectively. The threshold of 50% was chosen, in part, because the 1994 reauthorization of ESEA permitted schools with 50% or more of their students eligible for free or reduced price lunch to be declared Title 1 schools, allowing them to use Title I funds on a school-wide basis rather than having to direct them to individual students. Some researchers have used 75% as a threshold, but we found there were too few schools with percentages of eligible students above that level to yield results with acceptable levels of uncertainty.

The classification of a school as S1 or S2 was made on the basis of a NAEP school variable which identifies schools as being in one of eight categories, based on the reported percentage of students eligible for free or reduced price lunch. When that variable was missing, classification was made on the basis of a calculated ratio of the number of eligible students in the grade to the total number of students in the grade, both variables obtained from the CCD files. When missing data precluded calculating the ratio, we attempted to impute the correct classification based on the median income for the school (from the CCD file) and the relationship between median income and the lunch-program percentage for other schools with the same type of location and minority-percentage designation. The proportion of imputations for a data set varied from 0% to about 10%, with a median value of about 5%. In only a few cases were we unable to categorize a school and forced to exclude its data from the analyses.

Clearly, there is some uncertainty in the correct classification of a school in S1 or S2 (even when there are no missing data), particularly for those schools whose true values are near the boundary value of 50%. The effect of any misclassification would be to reduce the apparent differences between the two strata on any characteristic that was correlated with the S1/S2 classification.

The number of schools in the 30 NAEP state samples ranged from about 80 to about 100. Such numbers preclude very fine stratification of schools. After consideration of the focus of the study and the issues likely to be of greatest interest to policymakers (as well as some exploratory analysis), it was decided to retain only the S1/S2 stratification for each state/administration combination. Information on type of location and percentage minority was retained for later use in modeling.

It is important to note that while the definitions of S1 and S2 are the same across administrations, the universe of schools corresponding to each stratum did change over the period. In addition to school openings and closings, the lunch-program percentage in a school could have changed enough over the eight years from 1992 to 2000 to cause the school to cross the 50% boundary. Nonetheless, statistics based on properly weighted student scores from the schools in the

[12] The Common Core of Data is an annually updated database compiled by NCES that includes, among other things, information about schools that is not available through NAEP.

NAEP sample that fall in a particular stratum (in a specific year) yield approximately unbiased estimates of the corresponding population characteristics of the stratum (for that year).

With this structure we committed ourselves to carrying out a basic analysis for each combination of year by state by stratum, or 60 (= 3 x 10 x 2) analyses in all. Although this is a substantial number, it does give us the capacity to compare the time trends in the two strata within a state both with each other and with that of the state as a whole. It also provides us with the opportunity to contrast the record of a particular stratum within a state with those of equivalent strata in other states. This should yield more meaningful comparisons, giving us a more accurate picture of the degree of success a state has had in improving achievement over all, as well as in reducing the gap between Black students and White students. (Of course, since these analyses are purely descriptive, no causal mechanisms are offered or should be inferred.)

## *Descriptive Analyses*

Our program of quantitative analysis falls naturally into two phases. The first is more exploratory and the second is more model-based. In the first phase, our approach was to proceed systematically from higher to lower levels of aggregation. That is, we began with overall comparisons among states and between strata within states. These are not only informative but also provide a substantively meaningful background against which to judge the results of the more focal analyses, which track over time the achievement gap for each stratum within each state. In our presentation of the data, we have attempted both to identify general patterns and to highlight those states or strata whose results are sufficiently different to merit our attention.

All reported statistics are calculated using sampling weights so that they are approximately unbiased estimates of the corresponding population quantities. Each statistic is accompanied by an estimated standard error, obtained through standard NAEP procedures based on the jackknife method (Burke & James, 1997). When necessary, we employed specialized methods to take account of sampling dependencies. For example, the standard error of the mean difference between Black students and White students within a stratum requires such a calculation since these students are grouped by school, with many, if not most, NAEP school samples including students of both races.

## *Multilevel Modeling*

### *Rationale*

The second phase relies on hierarchical linear models or HLM (Raudenbush & Bryk 2002). This approach was motivated by the observation that the achievement gaps between Black and White students in each stratum in each state are generally large and persistent. However, in a particular stratum/state/year combination, Black students and White students are not distributed proportionately across schools. To the extent that overall school means are correlated with differences in the distributions of students of the two races, the estimate of the stratum-level achievement gap is confounded with differences in school means. One way to remove this confounding is to estimate a pooled within school achievement gap. That is, in effect, estimate the gap for each school in which students of both races are assessed and then compute an appropriate summary of those estimates. This procedure is easily handled by HLM.

Going a bit farther, it is natural to pose two related questions: First, how much is the estimated achievement gap reduced by taking into account student characteristics other than race?

Second, what proportion of the variance in school means can be accounted for by school characteristics? Both questions should be helpful in understanding the achievement gap. The first question recognizes that students differ with respect to a number of characteristics that are associated with achievement. Indicators of some of these characteristics are available through NAEP. By including them in our models, we can estimate how much of the pooled within school achievement gap is accounted for by these characteristics. The second question recognizes that certain school characteristics may account for differences in school means. Again, both questions, and their extensions, can be addressed through the development and estimation of an appropriate family of HLMs.

It should be noted that one unavoidable drawback to the HLM approach is that only schools with both Black and White students in the NAEP sample can contribute to the estimation of the achievement gap. For some stratum/state/year combinations, the number of such schools is rather small and leads to relatively large estimated standard errors for that contrast. However, the set of schools incorporated in the HLM analysis includes those schools with NAEP samples containing Black students but no White students or those with White students but no Black students. The latter two kinds of schools have information that can contribute to the estimation of the other regression coefficients in the model.

### *HLM and Plausible Values*

A NAEP assessment consists of a large number of items covering a broad domain. To limit testing time, only a small fraction of the item pool is administered to any one student. To accomplish this, the item pool is first organized into blocks of items. The blocks are then paired according to a balanced incomplete block or a partially balanced incomplete block design into booklets (Lazer, 1999). Booklets are randomly distributed to students within schools.

Because each student responds to relatively few items, a point estimate of an individual's score based on her cognitive data would be subject to substantial measurement error and, more importantly, would yield biased group estimates. Accordingly, through a rather complex process, NAEP estimates a latent proficiency distribution for each student, conditioning on the student's responses to the cognitive items, as well as to the background questions (Allen, Johnson, Mislevy, & Thomas, 1999). Five random draws are then made from the student's latent proficiency distribution. These are called plausible values.

The procedure for obtaining estimates of population quantities and the corresponding estimated standard errors consists of two steps: 1) Calculate the point estimate of the statistic. This is done by computing the statistic five times, once for each set of plausible values, and taking the average; and 2) Estimate the (total) variance of the statistic. This is done by computing an estimate of the variance due to the sampling of students and schools employing data from the first set of plausible values. A second component, representing measurement error, is obtained by computing the variance among the five point estimates calculated in step 1. The total variance is an appropriately weighted combination of the two components.

To facilitate applications to NAEP data, the HLM5 program was adapted for use with plausible values. This new program, HLM5-PV, automatically analyzes a NAEP dataset five times, once for each set of plausible values. (See Raudenbush, Bryk, Cheong, and Congdon (2001) for a full description of the HLM5 software.) The results displayed in the summary output of HLM5-PV reflect steps 1 and 2 above.

### *HLM and Sampling Weights*

The problem of whether and how to incorporate weights in fitting HLMs to survey data is an area of active research (Little 2003; Chambers 2003; Pfefferman, Skinner, Holmes, Goldstein, & Rasbash, 1998; Pfeffermann, Moura, & Nascimento Silva, 2004). As the discussion following the earlier Pfefferman paper indicates, there is no unanimity in the field with respect to this question, even as to whether weights should be used at all. Alternative suggestions are made, but there is no consensus on a preferred approach. The Pfefferman et. al. (2004) paper explores a promising strategy based on factoring the student design weight into two components: A component that is related to the selection of schools and another that is related to the selection of students within schools. One could then apply the weights to sample versions of certain census estimators. Unfortunately, the version of HLM that was available when the current analysis was done did not offer this option.

Upon reflection, we decided to forego entirely the use of weights. Our rationale was that, quite aside from the difficulty of implementation, only those schools in which both Black students and White students were assessed could contribute to the estimate of the achievement gap in a particular stratum/state combination. That set of schools did not constitute a simple random sample of the full set of schools for that stratum/state. Consequently, the weights we had available were not the appropriate weights and so, even were we to employ them in some way, we would not be able to make inferences to a larger population of schools.

Consequently, we fit the multilevel models without weights. This complicates comparisons with the descriptive analyses we have just outlined, which do use weights. Accordingly, we carry out some intermediate analyses that enable us to estimate how much of the differences between the sets of estimates may be due to differences in the weighting strategies. The intermediate analyses are described in the section containing the HLM results.

### *Centering Predictors*

There are several ways that level 1 predictors can be centered, e.g. no centering, centering around the grand mean of the full sample or centering around the group (i.e., school) mean. Centering the level 1 predictors affects the values and interpretations of the regression coefficients at both levels (Raudenbush & Bryk, 2002, Ch. 2). We chose to center each level 1 predictor around its corresponding school mean. Consequently, we can interpret the estimated Black/White gap as the pooled-within-school difference between Black and White students. When other predictors (also school-mean centered) are included in the level 1 model, the estimated Black-White gap is the pooled-within-school difference between Black and White students adjusted for these other student characteristics. When all predictors are school-mean centered, the level 1 intercept is the average outcome in the school.

Thus, when the level 1 intercept is treated as the criterion at level 2, the full between-school variation in school means is being modeled. Raudenbush and Bryk (2002, Ch. 5) present a similar centering strategy for estimating Black-White achievement gaps. When there is a structured regression at level 2, all level 2 predictors (i.e., school characteristics) are grand-mean centered. With this choice, the intercept in the regression for the level 1 intercept can be interpreted as the mean response of a school with the grand mean value on each level 2 predictor.

## *Model Building*

### *Foundations*

HLM can accommodate a large number of predictors at both level 1 (between students within schools) and level 2 (between schools). A fully specified HLM will include every level 1 parameter as an outcome at level 2. The resulting model can be difficult to interpret. It can be further complicated by classical regression problems, such as supressor effects and multicollinearity. In order to construct a meaningful and usable model, it is advisable to proceed by developing and testing simple models, which are systematically augmented and pruned until an acceptable final model is obtained (Raudenbush & Bryk, 2002). The rest of this subsection describes the process we carried out. Note that in each model all random errors are assumed to be mutually independent.

### *Unstructured Model*

The usual starting point is an unstructured model, in which there are no predictors at either level 1 or level 2. It has the following form:

Level 1
$(mathproficiency)_{ij} = B_{0j} + e_{ij}$, for student *i* in school *j*.
Level 2
$B_{0j} = \gamma_{00} + r_{0j}$
with var($e_{ij}$) = $\sigma^2$ and var($r_{0j}$) = $\tau_0^2$.

This is equivalent to an analysis of variance in which $\sigma^2$ is the variance within groups and $\tau_0^2$ is the variance between groups. This model is instructive because it tells us how much of the total variance in the outcomes is between schools. If there is none (or very little), a multi-level analysis is not needed. An estimate of the proportion of variance between groups, or the intraclass correlation, is calculated as:

$$\frac{\tau_0^2}{\sigma^2 + \tau_0^2}.$$

### *Structured level 1 model, unstructured level 2 model*

The next step is to introduce predictors at level 1, leaving level 2 unstructured. The predictors are represented by *X*'s, with the asterisks denoting that the predictors are school-mean centered. With this family of models, we can identify those student-level predictors that are statistically related to the outcome. This is exactly the kind of exploration that takes place in ordinary regression analysis. At this stage, we can also determine whether any of the level 1 regression parameters has a random component; i.e., whether a regression parameter varies substantially across schools. The model has the form:

Level 1

$(\textit{math proficiency})_{ij} = B_{0j} + B_{1j}X^{*}_{1ij} + ... + B_{pj}X^{*}_{pij} + e_{ij}$ , for student *i* in school *j*.

Level 2

$B_{0j} = \gamma_{00} + r_{0j}$

$B_{1j} = \gamma_{10} + r_{1j}$

•

•

•

$B_{pj} = \gamma_{p0} + r_{pj}$ ,

With var($e_{ij}$) = $\sigma^2$ and

$$\mathrm{T} = \begin{bmatrix} \tau_0^2 & \tau_{01} & \cdots & \tau_{0p} \\ \tau_{10} & \tau_1^2 & & \\ \vdots & & \ddots & \vdots \\ \tau_{p0} & \tau_{p1} & \cdots & \tau_p^2 \end{bmatrix}.$$

representing the variance-covariance matrix of the residuals in the level 2 model. By testing whether the variance terms, $\tau_0^2, \tau_2^2, \cdots, \tau_p^2$, are significantly different from zero, we can decide which of the regression coefficients $B_{0j}, B_{1j}, ..., , B_{pj}$ to treat as fixed and which as random. Specifically, if the variance term corresponding to a level 1 regression coefficient is nonzero, then we conclude that parameter varies over schools and should serve as a criterion in a between-schools regression.

The NAEP database makes available a large number of student covariates. These are described in Appendix E. Following a series of exploratory analyses, the initial pool of student covariates was reduced to three variables to be included in the level 1 model. Predictors were retained if the associated regression coefficient was of constant sign and statistically significant for most of the 60 analysis sets. The three covariates retained were: Student socioeconomic status (SES), student academic focus (AcadFoc), and the Black vs. White contrast (BvsW).

Briefly, SES is an index based on a combination of parental education and the number of reading related items in the home. AcadFoc is an index based on a combination of types and characteristics of math classes taken, student effort, and student beliefs. (Note: The components of AcadFoc vary across years.) BvsW is one for White students, 0 for Black students and missing for other race/ethnic groups. For further details, consult Appendix E.

Once the final level 1 model was obtained, the intercept and the three regression coefficients were tested to determine if the corresponding variance components were nonzero. Only the intercept term, reflecting the school mean, had a variance component significantly different from zero. In terms of the model parameters, the variance associated with the intercept, $\tau_0^2$, was found to be significantly different from zero, while the variances associated with the other regression terms were found to be *not* significantly different from zero. The inference is that typically, within a stratum, the school means were significantly different from one another, but the vectors of regression coefficient are the same across schools. (Of course, this finding substantially simplified the final set of analyses.) The model then takes the form:

Level 1

(*math proficiency*)$_{\text{ij}}$ = $B_{0j} + B_{1j}X^{*}_{1ij} + B_{2j}X^{*}_{2ij} + B_{3j}X^{*}_{3ij} + e_{ij}$, for student *i* in school *j*.

Level 2

$B_{0j} = \gamma_{00} + r_{0j}$,

with var($e_{ij}$) = $\sigma^2$, var($r_{0j}$) $= \tau_0^2$ and $B_{1j} = B_1, \ldots, B_{3j} = B_3$ for all *j*.

*Structured level 1 model, structured level 2 model*

Once the level 1 structure is finalized (i.e., predictors are chosen and it is determined which are fixed and which random), predictors for the random effects are introduced in level 2. In this study, predictors were retained if they were significant in most of the 60 analysis sets. The final model contained 4 predictors in the regression model with the school intercept from level 1 as the criterion. The predictors are Average School SES (AggSES), Percent Black Students Assessed (AggBvsW), School Climate, and Aggregated Academic Focus (AggAcadFoc). The predictors are represented by *W*'s, with the asterisks denoting that the predictors are grand-mean centered. Consult Appendix E for definitions. The final model has the following form:

Level 1

(*mathproficiency*)$_{\text{ij}}$ = $B_{0j} + B_{1j}X^{*}_{1ij} + B_{2j}X^{*}_{2ij} + B_{3j}X^{*}_{3ij} + e_{ij}$, for student *i* in school *j*.

Level 2

$B_{0j} = \gamma_{00} + \gamma_{01}W^{*}_{1j} + \gamma_{02}W^{*}_{2j} + \gamma_{03}W^{*}_{3j} + \gamma_{04}W^{*}_{4j} + r_{0j}$,

with var($e_{ij}$) = $\sigma^2$, var($r_{0j}$) $= \tau_0^2$ and $B_{1j} = B_1, \ldots, B_{3j} = B_3$.

*Defining trends for the analysis*

The full study analyzed data on the mathematics achievement of grade 8 students at three time points, 1992, 1996, and 2000. However, for the purpose of presenting findings on trends, only data from years 1992 and 2000 will be reported. After careful consideration, the authors decided that including the 1996 data would unduly complicate the summaries while contributing only marginally to the link with policy analysis.

When we report on the estimated Black-White achievement gaps from the HLM analyses, we present two sets of estimates. The first set contains what we term adjusted gaps because they are free of average differences in mean scores among schools. Specifically, in the model below, $B_1$ represents the average difference between Black and White students attending the same schools, pooled across schools in the stratum.[13] That is, $B_1$ is treated as fixed across schools. $B_{0j}$ represents the mean over all students in school *j* and is allowed to vary randomly over schools. The regression models are:

[13] Clearly, these estimates draw only on data from schools for which the NAEP sample included both Black and White students.

Level 1

(mathproficiency)$_{ij}$ = $B_{0j} + B_1 * BvsW^*_{ij} + e_{ij}$

for student *i* in school *j*, where, $BvsW^*_{ij}$ is the Black-White indicator deviated from its respective school mean.

Level 2 (no predictors)

$\mathrm{B}_{0\mathrm{j}} = \gamma_{00} + \mathrm{r}_{0\mathrm{j}}$

Table 1

*New York S2 (High Poverty Stratum). Pooled Within-School Black-White Achievement Gaps.*

| Model | 1992 | 2000 | 1992–2000 |
|---|---|---|---|
| Adjusted | 17.9 | 6.9 | 11.0 |
| Fully Adjusted | 14.2 | 5.1 | 9.1 |

Estimates of trends are simply differences of the estimates of $\mathrm{B}_1$ in 1992 and 2000. For example, see the *adjusted* results in Table 1 for S2 (the high-poverty stratum) in New York, which indicate that, over the period of interest, the achievement gap was reduced by 11 points.

The second set of estimates contains the *fully adjusted* gaps. These estimates are the pooled-within-schools estimated achievement gaps, adjusted for student differences in SES and Academic Focus and are also free of differences in mean achievement across schools. (Here, $\mathrm{B}_1$ is a partial regression coefficient with respect to the other predictors in the model.) While $\mathrm{B}_{0j}$, the average outcome in school j, is allowed to vary randomly over schools, the three regression coefficients are assumed to be constant across schools. Using data from the same reduced set of schools as before, the second set of estimates is derived by fitting the model below:

Level 1

(mathproficiency)$_{ij}$ = $B_{0j} + B_1 * SES^*_{ij} + B_2 * AcadFoc^*_{ij} + B_3 * BvsW^*_{ij} + e_{ij}$ for student *i* in school *j*, where the superscript '*' indicates predictors deviated from school means..

Level 2 (no predictors)

$B_{0j} = \gamma_{00} + r_{0j}$

The *fully adjusted* results in Table 1 indicate a reduction in the achievement gap of 9.1 points. To the extent that the *fully adjusted* gaps are smaller than the *adjusted* gaps in a particular year, it is possible to regard the covariates, SES and AcadFoc, as accounting for some of the observed differences in achievement between Black and White students. On average, we find that the *fully adjusted* gaps are smaller than the *adjusted* gaps by about 30 percent. Interpreting trends in *fully adjusted* gaps is difficult, however, since they are a function of differences between cohorts in both achievement and in the student characteristics employed as predictors.

## Findings

### *Education Policy*

All ten states embarked on some type of education reform during the period 1988 through 1998. Many of their efforts focused on setting academic standards for all students, adopting measures to improve teacher quality, developing new assessments of student academic achievement, and establishing accountability systems that were at least partially focused on student outcomes. Employing the policy framework we have developed, we now summarize our findings from a cross-state evaluation of the state profiles, expert interviews, and state syntheses. For each of the five policy levers in our framework, we identified a few key questions.

*Governance and the politics of reform.* What was the governance structure and politics around governance, in the context of reform? What was the balance between central and local control and did it change over the period? Who were the principal drivers of education reform? What were the main reform mechanisms?

*Education finance.* To what extent was there continuing commitment to improving the funding of education? What was the level and trajectory of the proportion of state funding in education? What was the level and trajectory of expenditures per student? What was the level and trajectory of the funding gap between high- and low-poverty districts?

*Curriculum and standards.* To what extent was there an ongoing commitment to improving curriculum and standards, especially in mathematics? Was there a strong state curriculum in mathematics (with statewide textbook adoption and alignment of textbooks with the curriculum)? Was the mathematics curriculum deep and rigorous?

*Teacher quality.* To what extent were there meaningful initiatives related to teacher quality? Was there a middle grades content-specific teacher certification, especially in mathematics? Was recertification tied to professional development? How well were teachers compensated compared to the nation as a whole and to the other states in the study? What was the extent of out-of-field teaching and how did it change over the period?

*Assessment and accountability.* Was there a broad commitment to assessment, especially of high-level skills? Was there continuity in assessment policy? Was there consistency in accountability policy? Was there a strong accountability system (with an effective system of sanctions and rewards) for both Title I and non-Title I schools?

#### *Analysis of Education Reform Policy Levers*

*Governance and the politics of reform.* The ten states varied substantially with respect to state control (see Table 2). For instance, California, New York, North Carolina, Texas, and Virginia all had strong state controls; whereas Maryland, Michigan, South Carolina, and Tennessee had strong local controls. Kentucky fell somewhere in-between, as authority was divided between the state department of education and the school districts.

Table 2
*Governance and Main Drivers of Reform*

| State | Strong State Control | Governor, State Legislatures | Court Case | Business Groups | Higher Education System | Grassroots & Special Interest Groups |
|---|---|---|---|---|---|---|
| California | X | X | | | X | X |
| Kentucky | X[1] | X | X[2] | X | | X |
| Maryland | | X | | | | |
| Michigan | | X | | X | | |
| New York | X | X | | | | |
| North Carolina | X | X | | X | | |
| South Carolina | | X | | | | |
| Tennessee | | X | X[3] | | | |
| Texas | X | X | X[4] | X | | |
| Virginia | X | X | | | | |

[1] Implementation of education reforms in Kentucky was divided between the state department of education and school districts. The state department of education created and managed the assessment and accountability systems, while the schools and school-based decision making councils decided on the curriculum and resource distribution.
[2] In 1989, the Kentucky Supreme Court ruled that the state's public school system was unconstitutional.
[3] In 1993, the Supreme Court of Tennessee ruled that the state's education finance system violated the its constitution's equal protection clause.
[4] In 1987, a Texas district judge ruled that the state's education finance system violated the state's 1984 Equal Opportunity Act, which aimed to reduce the monetary gap between funding for rich and poor school districts. In September 1991, the state began to comply by increasing spending in poorest districts.

As the entries in Table 2 indicate, in some states reform was driven "from the top," by the governor, the state board of education, or the superintendent of schools. These included Maryland, New York, South Carolina, and Virginia. In Kentucky, Michigan, North Carolina, and Texas, business groups also played a central role in providing impetus for reform. Moreover, Kentucky and California stand out as having had a wide variety of constituencies shaping education reform during the period.

For instance, the main drivers of Kentucky's educational reform during the late 1980s through the 1990s were a combination of local coalitions of grassroots non-governmental organizations, a coalition of school districts, a definitive state Supreme Court ruling, the governor, and the General Assembly, the state board and the commissioner of education. In California, the governor, some state legislators, and the superintendent of education were very active in the state's first wave of education reform from the mid 1980s through the mid 1990s. However, the state's higher education system, a large number of non-profit organizations and special interest groups also had a significant impact on the nature and the fate of the state's education initiatives. California also experienced considerable conflict with respect to both control of reform and the substance of the reforms (Wilson, 2003).

*Education Finance.* Clearly, funding is essential to the functioning of any education system. However, the impact of the funding is dependent on how it is allocated across the state, the activities it supports and the efficiency with which it is employed. Table 3 describes the ten states with respect to three indicators of state policies related to education finance: Average proportion of state funding, average expenditures per student, and funding disparities among school districts. As described in the table, the last indicator comprises two different statistics. The first is a measure of the funding disparities among districts in the state during the early part of the period of interest. (There is an implicit assumption that, generally, disparities favored lower poverty districts.) The second is a direct measure of the disparities between lower and higher poverty districts, compiled from data toward the latter part of the period of interest.

The table entries are the value of the indicator and a subjective ranking of the states into one of three categories.[14] A rank of 1 corresponds to the highest category and rank of 3 to the lowest category. We also constructed an overall ranking based on combining the three indicators, with the average level of expenditures per pupil having the greatest weight and the average proportion of state funding the least weight.

California was assigned rank 1 overall as it was the only state not in the lowest category for any indicator. Note that their per-pupil expenditures straddle the median for that indicator. At the other end of the scale, Virginia was the only state to be placed in the lowest category for three of the indicators. We did not find it possible to distinguish among the remaining eight states.

Interestingly, Kentucky, along with Tennessee and Virginia, employed education finance reform as one of their main policy levers during the period studied. In particular, Kentucky redesigned its school funding system to increase funding for students who required more time to achieve academic success and for teachers who could help these students succeed. Under the new funding formula, districts were required to meet certain local revenue-raising benchmarks; however, those with small tax bases and/or limited property values were entitled to additional state funding. As a result, Kentucky made some progress toward greater equity in spending per student.

Through Public Act 145 and Public Act 335, both enacted in 1993, Michigan completely restructured education funding. Local property tax was eliminated as the source of funding for the operating costs of K–12 public schools; instead, schools were funded through other sources, including a two-percentage point increase in sales tax and use tax, and the 50-cent increase in the cigarette tax. As a result, the state's share of education expenditures increased from about 30% to about 70% in just one year. PA 335 also increased support for professional development, at-risk students, and Math/Science Centers. It also extended the length of the school day from five to six hours.

In 1992, Tennessee overhauled its education finance system by creating the Basic Education Program (BEP), a regression-based formula that determined the funding level required for each school system to achieve a common and basic level of service for all students. Monies from BEP were allocated to both classroom and non-classroom components, including teachers' salaries, technology and other school improvements. After five years of graduated increases, full funding was attained in the 1997–98 school year.

[14] A state's ranking on an indicator is relative to the values of the other states on that indicator. We do not have absolute standards by which we can judge a state's record.

Table 3
*State Education Finance Indicators and Overall Ranking. Ranks in parentheses.*

| | | Average Funding Disparity Between High- and Low-Poverty Districts | | | |
|---|---|---|---|---|---|
| State | Average Proportion of State Funding[1] (percentage) | Gini[2] | Per-pupil gap[3] | Average Level of Expenditures[4] ($/Students) | Overall Ranking |
| California | 62.2 (1) | 0.082 (2) | $ 35 (1) | $ 5,858 (2) | 1 |
| Kentucky | 65.7 (1) | 0.092 (3) | -150 (1) | 5,576 (3) | 2 |
| Maryland | 38.2 (3) | 0.083 (2) | 701 (3) | 7,645 (1) | 2 |
| Michigan | 43.0 (2) | 0.100 (3) | 1,261 (3) | 7,690 (1) | 2 |
| New York | 40.6 (2) | 0.099 (3) | 2,794 (3) | 9,770 (1) | 2 |
| North Carolina | 65.2 (1) | 0.047 (1) | 413 (2) | 5,591 (3) | 2 |
| South Carolina | 49.5 (2) | 0.049 (1) | 427 (2) | 5,506 (3) | 2 |
| Tennessee | 45.7 (2) | 0.101 (3) | -138 (1) | 4,917 (3) | 2 |
| Texas | 42.2 (2) | 0.069 (1) | 386 (2) | 5,718 (2) | 2 |
| Virginia | 32.1 (3) | 0.100 (3) | 879 (3) | 6,470 (2) | 3 |

[1] Calculated by the National Center for Education Statistics (NCES) based on 1987–88 through 1997–98 data from the Common Core of Data (CCD) collection of surveys,

[2] Funding disparities are quantified by two methods. The first is the Gini coefficient, with higher values corresponding to greater disparities. Calculations are based on the Gini coefficients reported by Hussar & Sonnenberg (2000), which used the 1988 through 1990 financial data collected by the Census Bureau, as part of the Census Government F33 School District Finances Survey.

[3] The second funding-disparity measure is based on the ED Trust's (2002) calculation of gap between highest and lowest poverty districts 1997. The reported figures above are from Table 2 of the Education Trust (2002), *The funding gap: Low-income and minority students receive fewer dollars.* Washington, DC: Education Trust. The numbers read as follows: In 1997, the highest poverty districts in California received $35 less per student in state and local revenues than the lowest poverty districts. The highest poverty districts in Kentucky, on the other hand, received $150 more per student in state and local revenues than the lowest poverty districts.

[4] Constant prices at 2000–01, calculated by NCES based on 1992–93 through 1997–98 data from the Common Core of Data (CCD) collection of surveys.

*Curriculum and Standards.* Table 4 compares the extent to which the ten states had a continuous commitment to mathematics curriculum and standards. In particular, we examined the rigor and depth of the mathematics curriculum and associated standards during the period, as well as whether there was a statewide textbook adoption. All ten states had some type of mathematics learning standards in place by 1998. The standards evolved from being relatively vague to being more specific, and from focusing on basic level skills to focusing on higher order thinking skills. On the other hand, we found that states varied greatly in terms of rigor, depth, and quality of the mathematics curriculum and standards implemented, as well as the extent to which there was a statewide textbook adoption policy. The experience in California is of special note, as there were deep disagreements about the mathematics curriculum, which influenced assessment policy and other reform efforts.

Table 4

*Curriculum and Standards*

| State | Mathematics curriculum or standards | High quality curriculum or standards | Statewide textbook adoption in mathematics |
|---|---|---|---|
| California | X | X[1] | X[2] |
| Kentucky | X[3] | X[4] | |
| Maryland | X[5] | X[6] | |
| Michigan | X[7] | | |
| New York | X[8] | X[9] | |
| North Carolina | X | X | X |
| South Carolina | X | X[10] | X[11] |
| Tennessee | X | X[12] | X[13] |
| Texas | X[14] | X[15] | |
| Virginia | X[16] | X | X |

[1] California's state curriculum standards were considered to be high quality when first adopted but did not keep pace with the changing standards in math education. By 1996, they were considered only to be of adequate quality.

[2] California had statewide textbook adoption for grades K–8.

[3] The Kentucky content standards went through a process of evolution and developed their core content standards in 1998, at the end of the period being studied.

[4] Kentucky has a list of approved textbooks but schools make their own choices.

[5] The Maryland School Performance Assessment Program (MSPAP) shaped Maryland's curriculum to a great extent.

[6] The MSPAP became the *de facto* set of standards and MSPAP was generally considered to represent high standards.

[7] Michigan did not make its state standards mandatory for school districts.

[8] New York created state syllabi for every grade level and subject area. At the high school level, there are two sets of standards, leading to two types of diplomas, Regents and local.

[9] New York's high school math curriculum leading to the Regents diploma was broad, deep, and rigorous. Curriculum for the local diploma was much less demanding.

[10] Until 1993, South Carolina had very low-level math standards based on its basic skills tests. In 1993, math standards were completely re-written to meet with the national standards promoted by the National Council of Teachers of Mathematics. Grade-by-grade achievement standards were then created and implemented in 1998.

[11] South Carolina provided districts with a list of approved textbooks; however, local districts were able to select other texts and add to this eligible list of books. It was not until 1998 that the textbook list was aligned with the newly enhanced frameworks.

[12] Tennessee had strong and clear content standards into the mid-1990s but then shifted to a more locally controlled set of standards.

[13] Tennessee has statewide textbook adoption for more than 20 years. Although texts are nominally supposed to cover state standards, variation in both coverage and quality has been great, with a large number of texts being approved.

[14] Comprehensive statewide standards were only adopted in Texas near the end of our period of study.

[15] The state recommends (and subsidizes) but does not require textbook book adoption in all subject areas. During the 1988–98 period, there were not sufficient state standards to make textbook selections based on alignment.

[16] Since 1995, Virginia has demonstrated a strong commitment to a set of high quality learning standards.

In our view, the ten states fell into one of three categories during the period of interest. One state had a consistent commitment to curriculum and standards with a strong state mathematics curriculum: North Carolina. Other states had some commitment to curriculum and standards. A few of these states had a statewide textbook adoption policy that allowed either an extensive list of state approved mathematics textbooks, or substantial flexibility for local adoption that resulted in great variations in the quality and content of textbooks being used across school districts. Others demonstrated their commitment to standards only towards the end of the period being studied. These states are California, New York, South Carolina, Tennessee, and Virginia. The other states had inconsistent or weak commitments toward standards and curriculum: Kentucky, Maryland, Michigan, and Texas.

*Teacher Quality.* Table 5 compares states on various dimensions of teacher quality initiatives during the period. We were especially interested in how the states differed in terms of their middle grade math certification requirements, the amount of professional development required for re-certification, the extent to which the states established professional development centers and standards, the degree of out-of-field teaching in core subject areas in grades 7 to 12, and teacher salary levels. We found that only some states had a continuous and comprehensive commitment to teacher quality during the period.

Table 5
*Comparison of State Commitment to Teacher Quality*

| State | Middle grades math certification | Professional development tied to recertification every 5 to 7 years | State professional development centers/ standards | Extent of out-of-field teaching in core subject areas for teachers in grades 7–12 | Salary level |
|---|---|---|---|---|---|
| California | | X | X | Middle | High |
| Kentucky | | X | | High | Middle |
| Maryland | | | | Middle | High |
| Michigan | | X | | Low | High |
| New York | X | | X | Low | High |
| North Carolina | X[1] | X | X | Low | Low |
| South Carolina | X | X | X | Middle | Low |
| Tennessee | X[2] | | | High | Middle |
| Texas | | | X | High | Low |
| Virginia | X[3] | X | | Middle | Middle |

[1] Available since 1996.
[2] Available since 1997.
[3] Available but voluntary.

We grouped the states again in three categories. The states with continuous commitment to teacher quality: New York, North Carolina, and South Carolina. States with inconsistent commitment to teacher quality: California, Kentucky, Michigan, and Texas. States with weak or no commitment to teacher quality initiatives: Maryland, Tennessee, and Virginia.

More specifically, North Carolina, after 1996, required all new teachers to participate in a two-year mentoring and evaluation program; furthermore, middle school teachers were required to be licensed in a particular subject area, resulting in the state's having the lowest percent of out-of-field teaching in core subject areas among the ten states. During the study period, New York maintained a network of teacher centers to support professional development in localities across the state. In addition, beginning in the 1980's, middle grade teachers had to achieve subject-specific certification. This is reflected in the relatively low percent of out-of-field teaching in core subject areas found in New York. South Carolina supported a host of teacher improvement activities both statewide and targeted at low-performing schools. Although the state offered middle school certification, it was not required. This may explain why South Carolina ranks near the middle of the group of ten states with respect to out-of-field teaching in core subject areas.

California had a number of initiatives to improve teacher quality, both in terms of developing the skills of practicing teachers and enhancing the preparation of teachers coming into the system. However, with the class size reduction program and the steady population growth in the state over the period, increasing numbers of students were being taught by teachers with emergency certification or who were not certified for that subject. Kentucky revised its less rigorous standards for middle grade certification in 2000, when the state began to require new and re-certified teachers to pass a national teaching exam and to have two field specialties. However, the lateness of this initiative may explain why Kentucky consistently had the second highest proportion of classes with out-of-field teachers among the ten states. Michigan maintained rigorous teacher certification and re-certification programs, combined with high teacher salaries. The state has nearly the lowest percentage of out-of-field teaching among the states studied. Texas had invested in small-scale regional centers and university-based outreach programs. The state maintained its teacher certification program, which grants a permanent teaching certificate after the candidate has completed 30 semester hours of graduate coursework. As the state does not require specific middle grades certification, Texas, with a steady growth in the school-age population, has one of the highest percentages of out-of-field teaching.

Of the last group of states, Maryland, except for the granting of teacher certifications, essentially relegated responsibility for teacher standards and evaluations to the school districts and schools. Virginia, on the other hand, has had sporadic efforts to improve teacher quality but has largely left this effort to local school districts. Furthermore, the state offered an endorsement for middle school math instruction. Although the state also demanded the accumulation of professional development points for re-certification, points could be accrued through a wide-range of activities that could be certified by the district. Since 1994, Tennessee required all new teachers to have an academic major, a full semester of student teaching, and a strong general education core. For existing teachers, however, the state commitment to teacher professional development was minimal and varied with the amount of money available in the state budget. Middle grade certification was created in 1997 but was not content specific. Tennessee had the highest rate of out-of-field teaching of any state in this study. Tennessee also implemented a value-added model (Sanders & Horn, 1997) to evaluate teacher effectiveness. The use of the results was not mandated and less than half of the districts employed it in a serious and consistent fashion.

*Assessment and Accountability.* Table 6 presents a comparison of the ten states on key aspects of their assessment and accountability policies. States were examined with respect to the consistency of their commitment to assessment, the type of tests that was used, their ongoing commitment to accountability, and the effectiveness of their system of rewards and sanctions. Not surprisingly, we found that all the states had policies on assessment and accountability but with varying degrees of focus and with varying effectiveness in holding districts, schools, teachers, and students accountable.

Table 6
*Commitment to Assessment and Accountability*

| State | Consistent commitment to assessment | Assess higher-order skills | Consistent commitment to accountability | High-stakes accountability system of sanctions and rewards |
|---|---|---|---|---|
| California | X[1] | X[2] | | |
| Kentucky | X[1] | X | X | X[3] |
| Maryland | X[1] | X | X | |
| Michigan | X | X[4] | | |
| New York | X[5] | X | X | |
| North Carolina | X[6] | X | X | X |
| South Carolina | X | | X | |
| Tennessee | X[7] | | | |
| Texas | X | | X | X |
| Virginia | X[1] | | | |

[1] The type and form of assessment changed during the years of interest.

[2] California changed from relying on a more open-ended assessment that was aligned with relatively advanced learning objectives in 1993–1994 to norm-referenced standardized tests in 1995–1998.

[3] Kentucky's system of sanctions and rewards has evolved over time with increasing rewards and sanctions for schools and teachers.

[4] The content of the Michigan assessment evolved from basic skills to higher-level skills during the years of interest.

[5] New York maintained consistent commitment in assessing middle and high school students, but not elementary school students.

[6] North Carolina has consistently assessed elementary and middle school students since the late 1970s, but only consistently assessed high school students after 1996.

[7] Tennessee had a consistent assessment plan in the elementary and middle grades between 1992 and 1998, using a norm-referenced test in grades three through eight.

Three states had continuous and consistent commitment in assessment and accountability, with a strong accountability system incorporating sanctions and rewards: Kentucky, North Carolina, and Texas. Five states had continuous commitment to higher-level skills assessment, but had low-stakes accountability systems: California, Maryland, Michigan, New York, and South Carolina. The two remaining states, Tennessee and Virginia, had consistent commitment to basic skills assessment but had weak accountability systems.

More specifically, Kentucky demonstrated a clear commitment to assessment and accountability during the period of interest, although the mode of assessment, the level of accountability, and the system of rewards and sanctions evolved over time. Since the passage of the Kentucky Education Reform Act (KERA) in 1990, the mode of the standards-based assessment has changed from a norm-referenced to a performance-based and open-response portfolio system. All students in grades 4/5, 7/8, and 11/12 were assessed during this time. The state had a consistent commitment to accountability, although the level of focus shifted from the district to the schools and teachers. From 1988 to 1992, the system focused on accountability for whole school districts. From 1993 to 1998, accountability was much more centered at the school level. Furthermore, the rigor of the accountability system increased during the period of interest. For example, rewards and sanctions for teachers were imposed beginning in 1993 or 1994.

North Carolina had a consistent focus on assessment and accountability throughout the period of interest. The state had used some form of a statewide math assessment since the late 1970s in the elementary and middle grades. During the ten years studied here, the state used a multiple-choice exam that was aligned with state standards. The state also showed consistent commitment to accountability, focusing first on district level accountability. Beginning in 1996, it then focused on school and student level accountability with the introduction of the School Based Management and Accountability Program (commonly known as the ABCs). Furthermore, the state redirected its attention from both input and output measures to only performance-based measures after 1996. It imposed strong accountability on both Title I and non-Title I schools, with a system that used student progress (based on test scores) to award teacher bonuses, as well as school recognition or sanctions. For schools that were severely under-performing, the state had takeover teams of administrators and master teachers who worked for a period of a year to improve practice and performance.

Texas maintained a continuous effort in assessing students in basic skills and held schools accountable during the period. Since 1993–94, the state has demonstrated a consistent commitment to performance-based accountability by sending the message that improved student performance on the state test (TAAS) was the highest priority for schools and districts. Texas had the same system of accountability for both Title I and non-Title I schools. Sanctions on under-performing schools ranged from public reporting to reconstitution. In addition, student accountability came with the required exit exam. However, as the pressure to produce higher test scores intensified, many low-performing schools began to teach to the test, stage pep rallies, and conduct cram sessions and mock tests in the hope of raising scores. At higher-performing schools, other measures such as the SAT, AP tests, or in-school performance measures continued to define performance. TAAS performance remained high and of relatively little concern at the higher-performing schools.

Of the remaining states, five (California, Maryland, Michigan, New York, and South Carolina) had continuous commitment in accountability and assessment, although the mode or the type of assessment changed over time and the sanctions and rewards were not well articulated or implemented. The remaining two states (Tennessee and Virginia) had inconsistent commitment to assessment and almost no accountability system in place during the period of interest.

California was a leader in the drive for aligning assessment with curriculum and, in mathematics, structuring assessment to tap higher-order skills. The California Learning Assessment System (CLAS) was introduced in 1993 and was a model of a progressive assessment that employed both multiple-choice items and open-ended tasks. CLAS immediately ran into difficulties due in part to its novelty and in part due to its inability to meet all the goals set for it. Political battles around the curriculum, as well as a somewhat unfavorable review by a technical advisory committee, led to the early demise of CLAS. The state then shifted back toward more traditional norm-referenced tests and has continued to tinker with both the grades tested and the assessments used. Accountability during this period generally focused on public reporting of school results, and there was little in the way of rewards and sanctions.

Maryland continuously assessed students during the period of interest. The form of this assessment changed dramatically in 1991, when the state introduced the Maryland School Performance Assessment Program (MSPAP). The MSPAP was a performance-based test that assessed students in grades 3, 5, and 8 in math and English. The test did not yield individual scores and was used to measure class and school performance. The state used MSPAP for measuring both Title I and non-Title I schools' progress and reported achievement status. It appears, however, that meaningful sanctions were not imposed on low-performing schools, though there was some indication that teachers at the school level would get together and talk about how to improve student test performance.

Michigan has shown a continuous commitment to assessment over the past 30 years. Test content improved markedly in 1990, when the tests evolved into an assessment of higher level essential skills needed for academic and professional advancement. Michigan tested students in math and reading in grades 4, 7, and 11; and in writing, social studies and science in grades 5, 8, and 11. With the exception of the writing test, the assessments were all criterion-referenced, multiple-choice exams. Michigan maintained a very low-stakes accountability system. Although the state depended heavily on public reporting and school choice, it appears that this has had only limited influence on changing teacher practice. Michigan had a very weak accountability system for non-Title I schools and a slightly more rigorous system for Title I schools. Neither system resulted in significant sanctions for schools. Students and districts had almost no accountability at the state level during this time period.

New York has a long history of assessment, starting with the end-of-course Regents exams that were administered since 1878 to high school seniors. Over the years, the state developed several other exams to assess elementary and middle school students and high school graduates not taking the Regents exam. For instance, the norm-referenced Pupil Evaluation Program (PEP) was developed in 1965. It assessed reading and mathematics in grades 3 and 6 and writing in grade 5. PEP underwent several revisions, with the most significant in the early 1980s when standards were raised and the assessment became a criterion-referenced exam instead of a norm-referenced exam. In 1979, the state introduced a basic skills test, the Regents Competency Testing Program, in reading, writing, and mathematics for all high school students not taking the Regents exams in those areas. Although the state maintained some level of public accountability through reporting of test results, the state did not have strong accountability for its schools, with minimal sanctions or rewards. The accountability system was the same for Title I and non-Title I schools.

South Carolina used the same assessment system and maintained a consistent state accountability system focused on districts achievement in both cognitive and non-cognitive measures during the entire time period of interest in this study. The tests emphasized basic skills. The accountability system had limited effectiveness because the standards set represented a relatively low level of achievement, which allowed many districts to escape being identified as in need of assistance despite poor test results and high dropout rates.

Tennessee had a continuous commitment to assessment between 1992 and 1998 but accountability was minimal during the time period of interest. The state adopted the Tennessee Value Added Assessment System in the early 1990s and required all schools to participate with the passage of the Education Improvement Act legislation in 1992. Between 1992 and 1998, the state used norm-referenced tests in grades 3 through 8 to assess student progress. In addition, students had to pass an eighth-grade level competency test in order to graduate. However, with the exception of the relatively low-level graduation requirement, there were no real rewards or sanctions based on performance; furthermore, the state did not disaggregate data until the introduction of the No Child Left Behind Act.

Although it continuously assessed students at the elementary, middle, and high school levels, Virginia had a number of different assessment mechanisms during the time period under consideration. Through the 1980's and into the early 1990's, the state tested students in grades 4, 8, and 11 using nationally-normed achievement tests. These assessments were used to chart progress but had no consequences for students or schools. Over that same period, the state also had a minimum competency test that was first administered to students in tenth grade. Passing this test was required for graduation and those who did not pass the first time were able to retake the test in eleventh and twelfth grades. Subsequently, the state went through a period in which it administered, in the sixth grade, a set of tests for basic skills in reading, writing, and math. In 1998, the state shifted to assessments aligned with the Standards of Learning. Furthermore, until 1998, Virginia had essentially no performance-based accountability; that is, school and district accreditation was based entirely on "inputs," a measure of the quality of staff and services being provided to students. The state began to consider performance-based accountability only after the learning standards had been implemented. Thus, only in 1998 was school accreditation linked to student performance.

### *Conclusion*

Clearly, state responses to calls for education reform have been far from uniform. They vary in the policies they employ, what roles are assumed by different parts of the state government and local education agencies, who is held accountable for student learning, and how consistently and coherently they carried out their educational initiatives over the period 1988 through 1998. This heterogeneity is a function of differences in history, political culture, educational governance structure and policies, state demographics, and educational performance. State responses also reflect how effectively state leaders were able to emphasize the same educational issues over time. We found that it was rare for any state to focus on a particular policy lever over the entire ten-year period. Instead, it was more likely for a state to focus on one or two policy levers for several years before shifting to other policy levers. Such shifts during the period studied makes categorizing the coherence, consistency, and quality of state policies a challenging task. However, we attempted to do so in the belief that these qualities are essential to understanding how state-driven reform plays out in schools and classrooms. Our results are presented below.

### *Summary rankings*

We now propose to categorize the ten states with respect to their strength in each of the policy components, as well as provide an overall ranking. Categorizations were based on reviews of all the information available about each state (i.e., state profile, state interviews, state synthesis, and state summary). Nonetheless, such an endeavor is fraught with difficulty, as well as involving an element of subjectivity. Furthermore, the "grading" of the states is essentially normative. We have no absolute basis for judging states' success in implementing education reform policies.

Thus, we consider each state in light of the results for the other nine states, and decide on (what we judge as) a fair placement in one of three categories. In *governance and politics of reform*, we evaluated the state in terms of whether the state had strong central control, whether the drivers of reform came from multiple sources, and whether multiple policy levers were used to initiate reform during the period of interest. In *education finance*, we identified the states that achieved higher levels of per pupil expenditures, greater funding equity across districts, and a greater proportion of state support for education. In *curriculum and standards*, we determined whether the state had a strong commitment in aligning mathematics curriculum with mathematics learning standards and whether there was statewide textbook adoption during the period of interest. In *teacher quality*, we analyzed the extent to which the particular state was committed to raising teacher quality by increasing the licensure requirements and/or providing extensive professional development in the area of mathematics teaching. In *assessment and accountability*, we examined the extent to which the state consistently assessed their students, the extent to which the assessment tool measured higher-order thinking skills, and whether the state was committed to making their districts and schools accountable using a high-stakes accountability system of sanctions and rewards.

Furthermore, the overall ranking is strongly influenced by the focal question our study is intended to address. That question may be framed as follows:

> Given the character and quality of a state's policy efforts during the period 1988–1998, in comparison to those of the other states, is it reasonable to expect that it

> would achieve relatively greater progress in closing the achievement gap—or increasing the scores of Black students—in comparison to those other states?

In this regard, we gave greater weight to the last three components, with the rationale that they can be expected to be more proximal determinants of classroom behaviors and efforts. Thus, a state was assigned to the highest category if, over the period in question and in relation to the other states, changes in policies and/or improvements in policy implementation would lead one to expect substantially greater improvement in test outcomes. Thus, were there a state that had a consistent, coherent, and broad-based policy effort already in place by the late 1980s and kept it in place through the 1990s, we would not necessarily expect great relative improvement—although we might anticipate high absolute scores. More realistically, if a state only achieved coherence and consistency in its efforts by the mid-1990s, then we might expect only "average" improvement over 1992 to 2000, even if the policies and their implementation were exemplary.

Table 7
*Ranking of states across policy components.*

| Ranking | Governance and politics of reform | Education finance | Curriculum and standards | Teacher quality | Assessment and accountability | **Overall** |
|---|---|---|---|---|---|---|
| 1 | KY<br>NC<br>TX | CA | NC | NY<br>NC<br>SC | KY<br>NC<br>TX | NC |
| 2 | CA<br>NY<br>VA | KY<br>MD<br>MI<br>NY<br>NC<br>SC<br>TN<br>TX | CA<br>NY<br>SC<br>TN<br>VA | CA<br>KY<br>MI<br>TX | CA<br>MD<br>MI<br>NY<br>SC | CA<br>KY<br>NY<br>SC<br>TX |
| 3 | MD<br>MI<br>SC<br>TN | VA | KY<br>MD<br>MI<br>TX | MD<br>TN<br>VA | TN<br>VA | MD<br>MI<br>TN<br>VA |

Table 7 presents the complete set of rankings. Considering the overall ranking, we note that: Only North Carolina was placed in the highest category. While it would have been preferable to have a more balanced allocation of states across categories, North Carolina's set of policy component rankings were substantially superior to those of any other state; Five states, California, Kentucky, New York, South Carolina, and Texas, fell in the second group. These states were committed to education reform but focused on only one or two policy levers. These states also allowed a certain degree of local choice that led to inconsistency across districts, and the remaining four states, Maryland, Michigan, Tennessee, and Virginia, fell in the third group. These states did not have strong state control and failed to align their policies to any significant extent during the period studied. Virginia, as already noted, only mounted a consistent effort in the middle of the period under study.

## Structure of the Achievement Gap

### *Preliminaries*

As we indicated earlier, it will prove useful to set the findings about the patterns in the Black-White achievement gap against a background of between state as well as other, within-state results. Accordingly, we begin our analysis with the data displayed in Table 8. The subtable at the bottom shows that in 1992 the average for the ten states was nearly four points below the national average, but by 2000 the differential was less than 2 points. In the main table, the ten states are listed in order of their mean achievement levels in 1992. The range is only ten points, from Virginia (268) to North Carolina (258). All the states experienced some improvement over the next eight years, with North Carolina showing the greatest increase (21.7) and California the least (1.3). However, these increases should be viewed in light of the changes in exclusion rates over the same period. These are presented in the last column and we note that North Carolina stands out with an increase of 10.6%. This has led some commentators to discount entirely North Carolina's improvement (Amrein & Berliner, 2002). For an alternative view see Braun (2004).

Table 8
*State NAEP results.*

| State | Achievement 1992 | Change in mean achievement (1992 to 2000) | Change in exclusion rate (1992 to 2000) |
|---|---|---|---|
| VA | **268** | 8.8 | 4.7 |
| MI | 267 | 11.1 | 5.9 |
| NY | 266 | 9.8 | 4.6 |
| MD | 265 | 11.8 | 5.9 |
| TX | 265 | 10.3 | 3.0 |
| KY | 262 | 9.3 | 4.9 |
| CA | 261 | **1.3** | 0.4 |
| SC | 261 | 5.6 | 0.9 |
| TN | 259 | 4.6 | **-0.3** |
| NC | **258** | **21.7** | **10.6** |

|  | Average Achievement 1992 | Average Achievement 2000 |
|---|---|---|
| Nation | 266.9 | 274.4 |
| 10 study states | 263.2 | 272.6 |
| Difference | 3.7 | 1.8 |

Since we are interested in examining patterns of achievement at a level below that of the state, in Table 9 we display counts of schools and students in the NAEP sample for each stratum/state for 1992 and 2000. For all ten states, in both years, the number of schools in the lower poverty stratum (S1) is always larger, and usually much larger, than the number of schools in the higher poverty stratum (S2). In addition, for all ten states, the number of schools in S2 in 2000 is greater than in 1992. The latter outcome is presumably due both to changes in the demographic profiles of the states and differential success in obtaining school participation in NAEP.

Table 9
*Number of schools and students by stratum and state.*

| | | 1992 | | | | 2000 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | # of Students | | | | # of Students | | |
| State | | # of Schools | Total | White | Black | # of Schools | Total | White | Black |
| CA | S1 | | 1903 | 1014 | 112 | 42 | 956 | 502 | 44 |
| | S2 | 22 | 613 | 122 | 64 | 29 | 672 | 100 | 88 |
| KY | S1 | 77 | 2067 | 1779 | 190 | 60 | 1458 | 1270 | 129 |
| | S2 | 27 | 689 | 604 | 55 | 37 | 836 | 677 | 118 |
| MD | S1 | 83 | 2186 | 1422 | 526 | 84 | 2004 | 1315 | 444 |
| | S2 | 10 | 213 | 21 | 173 | 20 | 397 | 34 | 307 |
| MI | S1 | 82 | 2167 | 1808 | 207 | 64 | 1560 | 1356 | 83 |
| | S2 | 19 | 449 | 106 | 290 | 21 | 415 | 133 | 235 |
| NY | S1 | 69 | 1783 | 1440 | 127 | 42 | 975 | 750 | 72 |
| | S2 | 16 | 375 | 54 | 173 | 31 | 658 | 101 | 280 |
| NC | S1 | 87 | 2334 | 1733 | 498 | 79 | 1804 | 1311 | 373 |
| | S2 | 16 | 435 | 157 | 239 | 25 | 550 | 178 | 294 |
| SC | S1 | 74 | 1909 | 1293 | 491 | 48 | 1179 | 831 | 282 |
| | S2 | 28 | 716 | 213 | 430 | 47 | 1127 | 449 | 607 |
| TN | S1 | 73 | 1966 | 1616 | 282 | 67 | 1654 | 1339 | 218 |
| | S2 | 21 | 519 | 254 | 231 | 28 | 578 | 285 | 255 |
| TX | S1 | 68 | 1771 | 1078 | 193 | 56 | 1313 | 782 | 150 |
| | S2 | 30 | 843 | 185 | 106 | 41 | 1004 | 281 | 153 |
| VA | S1 | 93 | 2470 | 1763 | 483 | 89 | 2145 | 1492 | 403 |
| | S2 | 10 | 240 | 115 | 103 | 16 | 324 | 84 | 192 |

The number of students in S1 ranges from 956 (CA/2000) to 2470 (VA/1992). The number of students in S2 ranges from 213 (MD/1992) to 1127 (SC/2000). Looking forward to the comparisons between White students and Black students, the counts for each group are generally respectable; there are, however, six instances in which the number of White or Black students falls below 100. Four occurred in S2 in 1992: CA (Black), KY (Black), MD (White) and NY (White); two occurred in S2 in 2000: MD (White) and VA (White).

Table 10 presents stratum means for 1992 and 2000, as well as the changes over that period. All strata in all states experienced gains. These ranged from 7 (NC, TN) to 24 (NC), with a median of 11 points in S1, and from 2 (CA) to 24 (NY) with a median of 11 points in S2. With the exception of California, Maryland, Tennessee, and Virginia in S2, all the stratum gains were statistically significant. Thus, NAEP results suggest there was real improvement in both strata for most of the states, but that the typical rate of improvement amounted to only about a point and a half per year.

Table 10
*Mean achievement by stratum and state (standard errors in parentheses).*

| | Lower poverty (S1) | | | | | | Higher poverty (S2) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1992 | | 2000 | | 2000–1992 | | 1992 | | 2000 | | 2000–1992 | |
| State | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. |
| California | 268 | (2.1) | 275 | (2.6) | 7 | (3.3) | 242 | (2.8) | 244 | (2.9) | 2 | (4.0) |
| Kentucky | 266 | (1.3) | 278 | (1.4) | 12 | (2.0) | 254 | (1.9) | 260 | (2.2) | 6 | (3.0) |
| Maryland | 268 | (1.3) | 284 | (1.4) | 16 | (1.9) | 229 | (5.6) | 238 | (3.5) | 9 | (6.6) |
| Michigan | 273 | (1.8) | 284 | (1.6) | 11 | (2.4) | 236 | (3.4) | 251 | (2.7) | 15 | (4.3) |
| New York | 277 | (1.5) | 288 | (1.5) | 11 | (2.2) | 235 | (5.4) | 259 | (4.5) | 24 | (7.0) |
| North Carolina | 261 | (1.2) | 285 | (1.3) | 24 | (1.7) | 243 | (3.3) | 264 | (3.0) | 21 | (4.4) |
| South Carolina | 266 | (1.2) | 274 | (2.1) | 8 | (2.4) | 245 | (2.0) | 259 | (1.8) | 14 | (2.7) |
| Tennessee | 263 | (1.2) | 270 | (2.0) | 7 | (2.4) | 242 | (2.6) | 247 | (3.4) | 4 | (4.3) |
| Texas | 271 | (1.7) | 283 | (2.0) | 11 | (2.7) | 251 | (1.5) | 266 | (2.1) | 15 | (2.6) |
| Virginia | 270 | (1.3) | 279 | (1.8) | 9 | (2.2) | 250 | (5.1) | 258 | (3.7) | 8 | (6.3) |
| Median | 268 | | 281 | | 11 | | 243 | | 259 | | 11 | |

In addition to evaluating the state-specific numerical outcomes, it can also be helpful to consider the stratum results in the aggregate. Figure 1 presents two back-to-back stem-and-leaf displays, for S1 and S2, respectively. The general improvement in both S1 and S2 over the period is quite evident. While this is welcome, comparisons between strata within a year are less heartening. In 1992, the distribution of means for S2 lies well below that of S1. Indeed, there is a seven-point gap between the state with the highest S2 mean (KY) and the state with the lowest S1 mean (NC). The story is essentially the same in 2000: There is a four-point gap between the state with the highest S2 mean (TX) and the state with the lowest S1 mean (TN).

Of greater concern, perhaps, is the comparison between the distributions for S1 in 1992 and S2 in 2000. We note that the highest-ranking states with respect to S2 in 2000 are just at the level achieved by the lowest-ranking states with respect to S1 in 1992. That is, overall in 2000, mean achievement in higher-poverty schools still fell well short of the levels attained by lower-poverty schools eight years earlier. This observation leads us to consider the gaps between strata on a state-by-state basis.

We first investigate, however, the possibility of a relationship between changes in stratum mean and the "baseline" level of achievement in 1992. Such a relationship could complicate the interpretation of the findings, since it raises the possibility that some of the observed patterns might simply be manifesting a statistical artifact such as regression to the mean. Figures 2 and 3 display the gains from 1992 to 2000 for each state for S1 and S2, respectively. The error bars attached to each state gain extend one standard error in each direction, corresponding to a 68% confidence interval. Leaving aside the outliers (NC in S1 and NY in S2), there does not appear to be even a modest relationship between gains and baseline.

| Lower Poverty Stratum (S1) | | | Higher Poverty Stratum (S2) | | |
|---|---|---|---|---|---|
| 1992 | | 2000 | 1992 | | 2000 |
| | 28* | 58 | | | |
| | 28 · | 344 | | | |
| 7 | 27* | 589 | | | |
| 310 | 27 · | 04 | | | |
| 8866 | 26* | | | 26* | 6 |
| 31 | 26 · | | | 26 · | 04 |
| | 25* | | | 25* | 899 |
| | 25 · | | 410 | 25 · | 1 |
| | | | 5 | 24* | 7 |
| | | | 322 | 24 · | 4 |
| | | | 65 | 23* | 8 |
| | | | | 23 · | |
| | | | 9 | 22* · | |

*Figure 1*. Stem-and-leaf plots for stratum means by state and year. (Note that the stems for the two displays are aligned numerically, so that comparisons can be easily made between the displays.) In each panel, the central column represents the "stem" and the side columns represent the "leaves." In 1992, for example, the state with the highest achieving S1 among the ten states had a score level of 277, while the lowest achieving S1 had a score level of 261. Note that the stems for the two panels are aligned vertically, facilitating comparisons across panels. For further guidance on reading stem-and-leaf plots, see Tukey (1977) or any modern introductory statistics textbook.

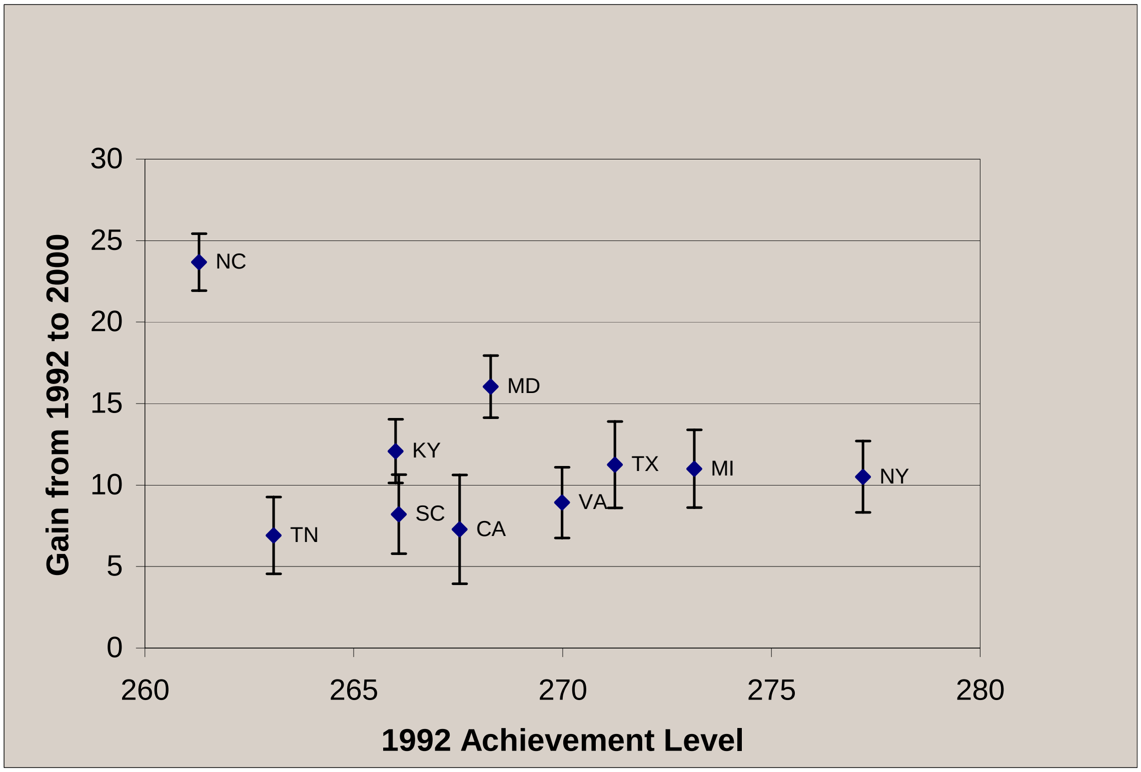

*Figure 2*. Lower poverty stratum (S1). Gain in mean achievement (1992 to 2000) vs. mean achievement in 1992.

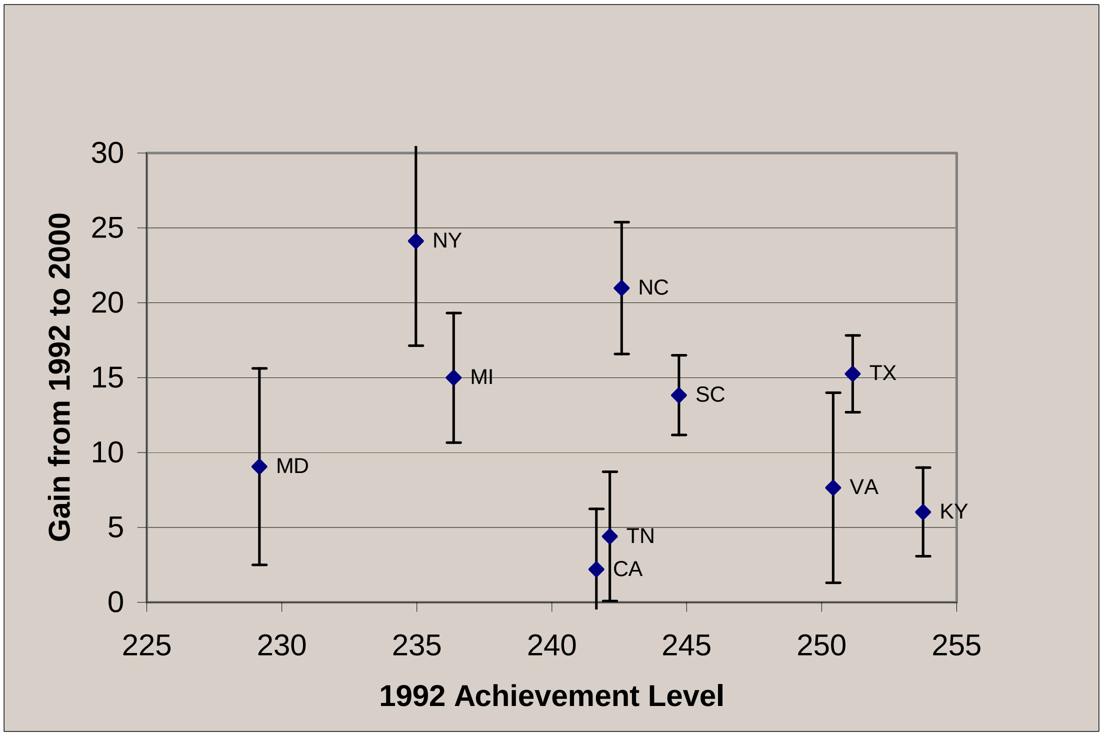

*Figure 3*. Higher poverty stratum (S2). Gain in mean achievement (1992 to 2000) vs. mean achievement in 1992.

## *Analysis of the Gaps*

Table 11 displays the stratum gaps for each state in both 1992 and 2000. The stratum gap is defined as the difference in mean achievement between S1 and S2. The states are listed in ascending order of the gap in 1992, with Kentucky having the smallest gap (12) and New York the largest (42). In 2000, Texas and South Carolina had the smallest gaps (16), while Maryland had the largest (46). Note that these stratum gaps are typically many times larger than the corresponding estimated standard errors.

The last panel in Table 11 displays the changes in the gaps from 1992 to 2000, with the corresponding standard errors. (Negative numbers signal increases in the gap and positive numbers signal decreases.) Since these changes are "differences of differences," the estimated standard errors are somewhat larger than before. Maryland, Kentucky, and California experienced the largest increases (-7, -6, -5), while New York and South Carolina experienced the largest decreases (+14, +6). Only New York's change is statistically significant. It is interesting to note that, despite having the largest increase in the stratum gap among the ten states, in 2000 Kentucky still had the third smallest stratum gap.[15]

Table 11
*Stratum gaps in mean achievement by state (standard errors in parentheses).*

| | [S1 - S2] | | [S1 - S2] | | Reduction in Gap[1] | |
|---|---|---|---|---|---|---|
| State | 1992 | | 2000 | | 1992 to 2000 | |
| KY | 12 | (2.3) | 18 | (2.7) | -6 | (3.5) |
| NC | 19 | (3.4) | 21 | (3.2) | -3 | (4.7) |
| VA | 20 | (5.3) | 21 | (4.1) | -1 | (6.7) |
| TX | 20 | (2.3) | 16 | (2.9) | 4 | (3.7) |
| TN | 21 | (2.9) | 23 | (4.0) | -3 | (4.9) |
| SC | 21 | (2.3) | 16 | (2.7) | 6 | (3.6) |
| CA | 26 | (3.4) | 31 | (3.9) | -5 | (5.2) |
| MI | 37 | (3.9) | 33 | (3.1) | 4 | (4.9) |
| MD | 39 | (5.7) | 46 | (3.8) | -7 | (6.8) |
| NY | 42 | (5.6) | 29 | (4.7) | 14 | (7.3) |
| Median | 21 | | 22 | | -2 | |

[1] Calculations for the last panel are based on original estimates of gaps in 1992 and 2000, rather than the rounded values presented in the table.

The distribution of gaps for the ten states, however, did not change very much over the period, with the median gap increasing slightly from 21 points to 22 points. It is noteworthy that in both years the range from the state with the lowest gap to the state with the largest gap was 30 points, or about three times the size of the overall difference between the highest and lowest ranking

[15] One can wonder whether legislators and other stakeholders in Kentucky should have been more concerned with losing ground in their effort to reduce the gap, or could take solace in maintaining their relative ranking. This contrast highlights the difficulty in making simple summaries of state achievement results over time.

states (see Table 8). This supports our contention of the importance of studying patterns of achievement within states. With these comparisons between strata as background, we now turn our attention to comparisons between Black students and White students.

## *Black-White Achievement Gaps*

Table 12 displays the mean scores for White students, Black students, the gaps (White mean—Black mean) for each state in 1992 and 2000, as well as the changes in the gaps over this period.[16] All the gaps displayed in Table 12 are many times larger than the corresponding standard errors. In 1992, Kentucky had the smallest gap (23) and New York the largest (47), yielding a range of 24 points. In 2000, Kentucky had the smallest gap (22) and Michigan the largest (44), yielding a range of 22 points. Equally important is the fact that the median of the state gaps was large and remained essentially constant over the period. The median gap in 1992 (34 points) is about 50% larger than the corresponding median stratum gap in 1992 (21 points), or about three times larger than the difference between the highest and lowest ranking states overall. There is considerable variability among states in the changes in the gaps, with North Carolina experiencing the largest increase (-7) and New York the largest decrease (+15). Both changes are statistically significant. In 2000, Kentucky was still the state with the smallest gap, while New York had moved to below the median.

Beginning our examination of within-stratum patterns, Table 13 displays the mean achievement for Blacks and Whites for each stratum/state/year combination, as well as the changes for both Blacks and Whites in each stratum/state from 1992 to 2000. There was an increase in mean achievement for both Black students and White students over the period for each stratum/state, although there was considerable variability among states. There were many cases of substantial increases for Black students. In S1, NY(23) and NC(17) had the best records; in S2, NY(29) and NC(21) again stand out. Both these states had excellent records with respect to gains for White students, although New York in S1 experienced a gain of only 9 points.

Inasmuch as a key goal of the study is to examine outcomes and trajectories in each stratum, Table 13 enables us to compare the gains over the period of White students in S1 to those of White students in S2. The gains in S1 were greater in five states and smaller in four states. However, the only state in which the difference in gains approached significance was Virginia. For Black students, the gains in S1 were greater in five states and lower in five states. In no state was the difference in gains significant.

[16] The estimated standard errors for the gaps take into account the dependency between the mean scores for White students and Black students induced by the clustering of students within schools.

Table 12
*Mean achievement for White students and Black students, and Black-White gaps, by state and year (standard errors in parentheses).*

| | 1992 | | | | | | 2000 | | | | | | Reduction in gap | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | White | | Black | | W-B | | White | | Black | | W-B | | 1992 to 2000 | |
| State | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. |
| CA | 277 | (1.9) | 234 | (3.6) | 42 | (3.9) | 278 | (2.2) | 242 | (2.8) | 36 | (3.6) | 6 | (5.3) |
| KY | 265 | (1.1) | 242 | (2.6) | 23 | (2.8) | 275 | (1.3) | 253 | (2.8) | 22 | (3.0) | 1 | (4.1) |
| MD | 279 | (1.5) | 240 | (2.0) | 39 | (2.4) | 290 | (1.3) | 249 | (2.0) | 41 | (2.2) | -2 | (3.3) |
| MI | 277 | (1.5) | 233 | (1.8) | 44 | (2.5) | 286 | (1.4) | 242 | (2.6) | 44 | (2.6) | 0 | (3.6) |
| NY | 280 | (1.1) | 233 | (4.4) | 47 | (4.5) | 289 | (1.3) | 257 | (4.3) | 32 | (4.4) | 15 | (6.3) |
| NC | 267 | (1.0) | 239 | (1.7) | 28 | (1.6) | 291 | (1.1) | 256 | (1.4) | 35 | (1.8) | -7 | (2.4) |
| SC | 274 | (1.1) | 242 | (1.0) | 32 | (1.3) | 279 | (1.5) | 249 | (1.7) | 30 | (1.8) | 2 | (2.2) |
| TN | 266 | (1.1) | 235 | (2.4) | 31 | (2.4) | 271 | (1.4) | 237 | (3.0) | 34 | (3.1) | -3 | (3.9) |
| TX | 279 | (1.5) | 244 | (2.0) | 35 | (2.3) | 288 | (1.4) | 252 | (3.3) | 36 | (3.0) | 0 | (3.8) |
| VA | 275 | (1.1) | 245 | (1.8) | 30 | (2.2) | 285 | (1.4) | 252 | (1.9) | 33 | (2.1) | -3 | (3.0) |
| Median | 276 | | 239 | | 34 | | 286 | | 251 | | 35 | | 0 | |

Table 13
*Mean achievement for Black and White students by stratum and state (standard errors in parentheses).*

| | 1992 | | | | 2000 | | | | 2000–1992 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | White | | Black | | White | | Black | | White | | Black | |
| State | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. |
| S1 (Lower poverty) | | | | | | | | | | | | |
| CA | 279 | (2.0) | 241 | (3.3) | 280 | (2.4) | 248 | (5.3) | 2 | (3.1) | 7 | (6.2) |
| KY | 268 | (1.3) | 244 | (3.1) | 281 | (1.4) | 258 | (3.0) | 13 | (2.0) | 14 | (4.3) |
| MD | 279 | (1.5) | 244 | (2.0) | 291 | (1.2) | 258 | (2.3) | 12 | (1.9) | 15 | (3.0) |
| MI | 278 | (1.5) | 241 | (2.8) | 287 | (1.5) | 249 | (6.0) | 9 | (2.1) | 8 | (6.6) |
| NY | 281 | (1.0) | 252 | (5.1) | 290 | (1.4) | 275 | (7.0) | 9 | (1.7) | 23 | (8.7) |
| NC | 267 | (1.0) | 242 | (2.0) | 293 | (1.2) | 259 | (1.8) | 25 | (1.6) | 17 | (2.7) |
| SC | 275 | (1.2) | 246 | (1.2) | 283 | (2.0) | 252 | (3.5) | 7 | (2.3) | 7 | (3.7) |
| TN | 268 | (1.3) | 241 | (2.3) | 274 | (1.7) | 246 | (6.1) | 6 | (2.1) | 5 | (6.5) |
| TX | 280 | (1.8) | 246 | (2.6) | 290 | (1.9) | 256 | (5.4) | 10 | (2.6) | 10 | (6.0) |
| VA | 276 | (1.1) | 247 | (1.6) | 285 | (1.5) | 255 | (2.4) | 9 | (1.8) | 8 | (2.9) |
| S2 (Higher poverty) | | | | | | | | | | | | |
| CA | 264 | (2.8) | 221 | (7.3) | 267 | (5.7) | 237 | (4.5) | 3 | (6.3) | 16 | (8.6) |
| KY | 257 | (1.9) | 236 | (4.9) | 263 | (2.2) | 247 | (4.5) | 6 | (2.9) | 11 | (6.6) |
| MD | 253 | (13.2) | 226 | (4.8) | 261 | (11.3) | 237 | (2.9) | 9 | (17.4) | 11 | (5.7) |
| MI | 266 | (5.1) | 227 | (1.6) | 280 | (4.7) | 240 | (2.7) | 14 | (6.9) | 13 | (3.1) |
| NY | 266 | (5.8) | 223 | (3.7) | 283 | (3.9) | 252 | (4.2) | 17 | (7.0) | 29 | (5.6) |
| NC | 262 | (3.5) | 231 | (2.1) | 280 | (3.1) | 252 | (2.2) | 18 | (4.7) | 21 | (3.1) |
| SC | 265 | (2.3) | 238 | (1.6) | 272 | (2.4) | 248 | (1.9) | 7 | (3.3) | 10 | (2.5) |
| TN | 257 | (2.0) | 227 | (3.0) | 261 | (3.2) | 230 | (3.8) | 4 | (3.8) | 2 | (4.8) |
| TX | 274 | (2.8) | 241 | (2.9) | 282 | (1.9) | 248 | (4.1) | 8 | (3.4) | 7 | (5.0) |
| VA | 260 | (3.8) | 240 | (7.2) | 279 | (4.5) | 247 | (3.5) | 19 | (5.9) | 7 | (8.0) |

We now turn our attention to Table 14, which displays the Black-White gaps for each stratum/state/year and the changes in the gaps over the period. This is the main focus of the section. The states are listed in descending order of the improvement (reduction) in the state level gaps, which are presented in the left-most column. In both 1992 and 2000, and for both S1 and S2, the median (for the ten states) within-stratum Black-White gap was only slightly smaller than the median (for the ten states) within-state Black-White gap (see Table 12). Indeed, that is the case for most states; i.e., for most states, the achievement gap within a stratum is almost as large as the achievement gap for the entire state. Thus, the state-level gap cannot be accounted for by the (substantial) gap between the strata and the differential distributions of Black students and White students across strata.

This finding is somewhat surprising, in view of the continuing discussions about the relationships between academic achievement, on the one hand, and race and class, on the other. Given the substantial gap in achievement between S1 and S2 in each state and the fact that Black students, in comparison to White students, are disproportionately enrolled in schools in S2, one might expect that the observed Black-White achievement gap at the state level was largely a consequence of Black students attending lower achieving schools. In fact, the magnitude of the state level gap was also due to the large differences in achievement between Black students and White students attending schools in the same stratum. Restricting attention to comparisons within stratum can only partially control for differences in other factors that are associated with achievement. In particular, students of different races are not proportionately distributed across schools within a stratum. As we indicated earlier, that observation motivated the HLM analyses that do control for individual school effects. The results of those analyses are presented in the next section.

Table 14
*Black-White gaps (White mean—Black mean) in mean achievement by stratum and state.*

| Reduction in State gap (1992–2000) | | | S1 (Lower poverty) gap | | | | | | S2 (Higher poverty) gap | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1992 | | 2000 | | 1992–2000 | | 1992 | | 2000 | | 1992–2000 | |
| Mean | S.E. | State | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. | Mean | S.E. |
| 15 | (6.3) | NY | 29 | (4.9) | 15 | (7.1) | 14 | (8.6) | 43 | (6.9) | 31 | (4.5) | 12 | (8.2) |
| 6 | (5.3) | CA | 37 | (3.6) | 32 | (5.7) | 5 | (6.7) | 43 | (7.9) | 30 | (6.7) | 13 | (10.4) |
| 2 | (2.2) | SC | 30 | (1.6) | 30 | (3.1) | -1 | (3.5) | 27 | (2.8) | 24 | (2.6) | 3 | (3.8) |
| 1 | (4.1) | KY | 24 | (3.3) | 23 | (3.4) | 1 | (4.7) | 21 | (5.3) | 16 | (4.6) | 5 | (7.0) |
| 0 | (3.6) | MI | 37 | (3.4) | 38 | (5.7) | -1 | (6.6) | 39 | (5.7) | 39 | (4.8) | -1 | (7.5) |
| 0 | (3.8) | TX | 34 | (2.8) | 34 | (4.5) | 0 | (5.3) | 33 | (4.5) | 34 | (3.8) | -1 | (5.9) |
| -2 | (3.3) | MD | 35 | (2.4) | 33 | (2.5) | 3 | (3.4) | 26 | (11.5) | 24 | (10.6) | 2 | (15.6) |
| -3 | (3.0) | VA | 29 | (2.0) | 31 | (2.5) | -2 | (3.2) | 20 | (6.7) | 32 | (6.1) | -11 | (9.1) |
| -3 | (3.9) | TN | 27 | (2.2) | 28 | (6.1) | -1 | (6.4) | 29 | (3.6) | 31 | (4.5) | -2 | (5.8) |
| -7 | (2.4) | NC | 25 | (1.8) | 34 | (2.1) | -9 | (2.7) | 31 | (3.7) | 28 | (3.5) | 3 | (5.1) |
| 0 | | Median | 29 | | 31 | | 0.5 | | 30 | | 30 | | 2.5 | |

With respect to reducing the Black-White gap in S1, NY(+14) and CA(+5) have the best records while NC(-9) has the poorest. Only the result for North Carolina reaches statistical significance. Turning our attention to S2, we see that CA(+13), NY(+12) have the best records, while VA(-11) has the poorest. None of the results reach statistical significance. The trends in Table 14 should be interpreted in light of those presented in Table 13. For example, there was a 17 point increase for Black students in S1 in North Carolina, but that was still smaller than the 25 point increase for White students in S1. The difference in gains is reflected in the poor gap trend result for the state. Virginia presents another interesting case. From Table 13, we note that Black students posted increases of 8 and 7 points in S1 and S2, respectively. Although White students in S1 did only a little better (9 points), those in S2 posted a larger gain of 19 points, leading to a 12 point increase in the gap. Overall, the median reduction in the achievement gap in S2 was 2.5 points.

Turning our attention to comparisons between strata (within states), we note that the reduction in the gap in S1 was greater than that in S2 for five states and lower in four states. However, the magnitude of the differences in gaps were generally small and approached significance only in one state, North Carolina.

It may be misleading to focus on changes in the Black-White gaps without taking into account the patterns of achievement for Black students and White students separately. Accordingly, we present Figures 4 and 5, which display scatterplots of the reduction in the achievement gap and the change in mean achievement for Black students in S1 and S2, respectively. (Again, the error bars extend one standard error in each direction, corresponding to a 68 percent confidence interval.) In both S1 and S2, New York is an outlier with respect to each dimension. There is no relationship apparent in S1 while, in S2, there is only a hint of a positive relationship.

Thus, in general, states that appear exemplary in the analysis of a particular educational outcome, may be viewed quite differently in another analysis. Leaving aside North Carolina for a moment (in view of the questions that have been raised about the change in exclusion rates), consider California, which appeared to make some progress in reducing the Black-White gaps within strata, while making little dent in the stratum gap. New York presents a more consistent picture, displaying typical to exemplary progress at each level of analysis.

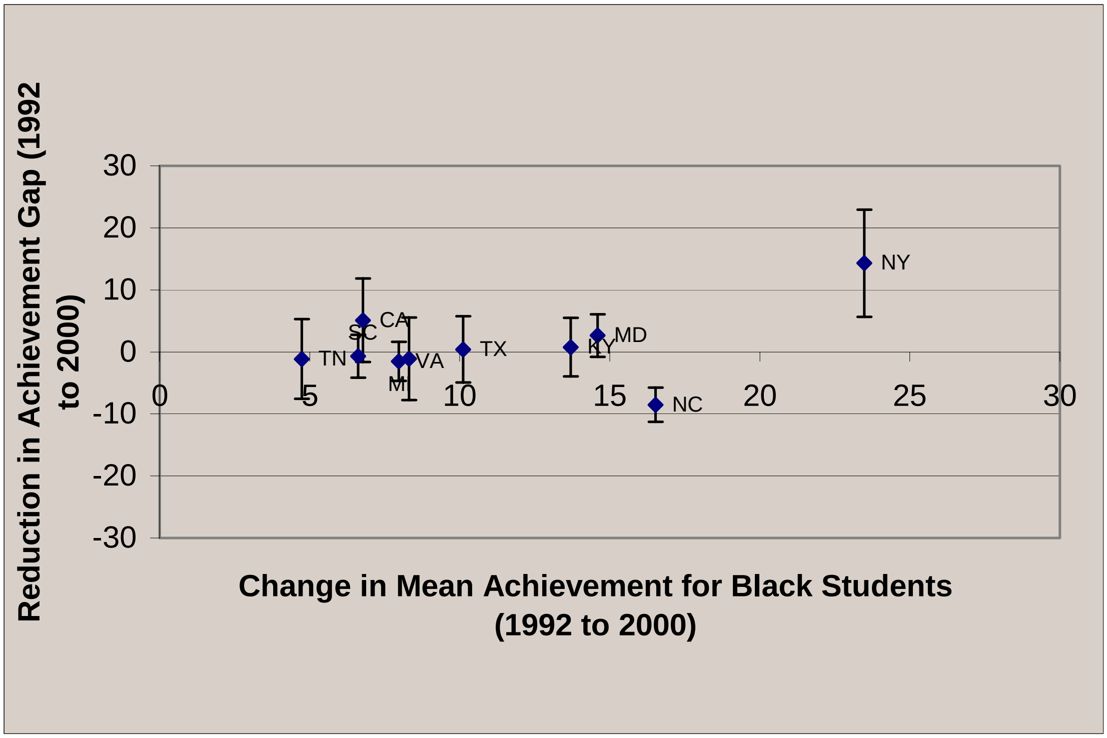

*Figure 4*. Lower poverty stratum (S1). Reduction in achievement gap (1992 to 2000) vs. change in mean achievement for Black students (1992 to 2000).

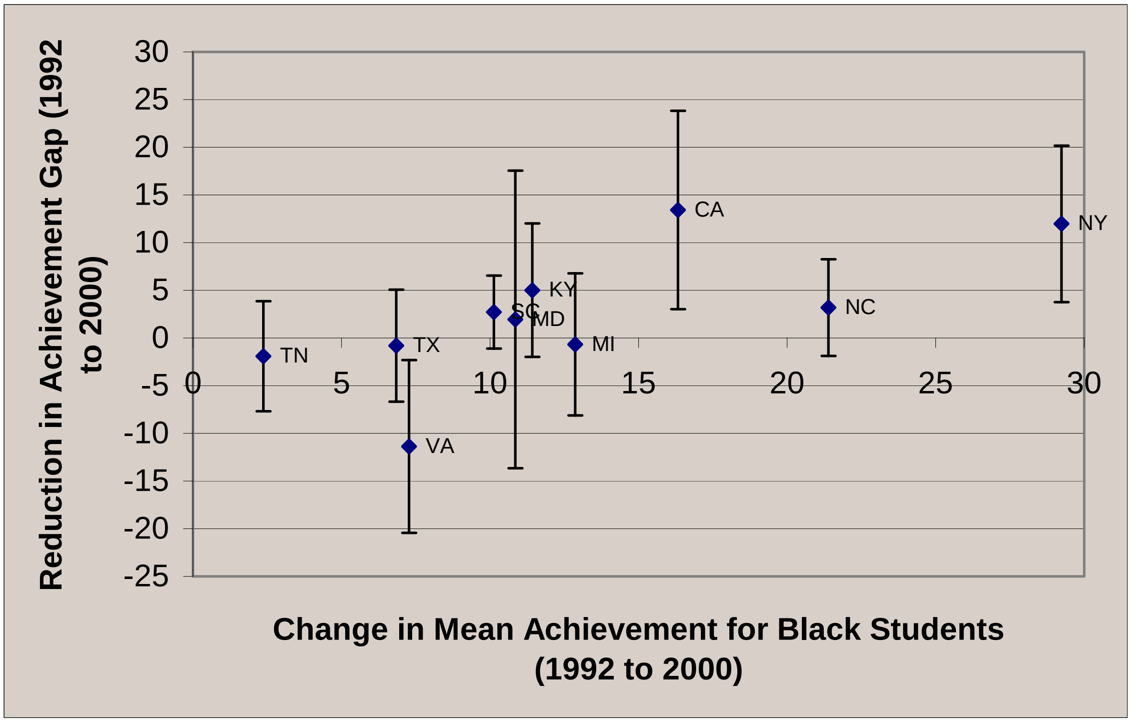

*Figure 5*. Higher poverty stratum (S2). Reduction in achievement gap (1992 to 2000) vs. change in mean achievement for Black students (1992 to 2000).

## *Grading the States*

The principal aims of the previous subsection were to present estimates of the Black-White achievement gaps at both the state and the stratum within state levels, as well as to describe the changes in the three sets of gaps over the period of study.[17] While we believe these descriptions are of interest in their own right, they can also serve as the basis for judgments about the (relative) success that the states have had in improving the outcomes for Black students. These judgments will be summarized in a "report card" for the states that can be juxtaposed against the education policy report card that was presented in the previous section. The comparison of the two report cards in the linking section will offer some insight into the relationship between state-initiated reforms and student outcomes.

There is no absolute basis for evaluating a state's success in raising test scores: The grades assigned to the states are normative, in that each state's record is considered in the context of the results for the other nine states. Although there is an element of subjectivity in such an effort, we believe that the summary provided by such a ranking serves a useful purpose in facilitating the linking of policy and outcomes.

The states have been ranked separately with respect to two outcomes in each stratum: improving the achievement of Black students and reducing the achievement gap. For the first outcome, we consulted Table 13, and for the second, Table 14. For each combination of stratum

[17] This program is continued in the following section, where HLMs are employed to estimate the achievement gaps within schools within a stratum.

and outcome, states were classified into one of three categories. In ranking a state, we considered both the absolute magnitude of the change in relation to those of other states, as well as the effect size of the change (i.e., the ratio of the change to its estimated standard error). The results are presented in Table 15.

Table 15
*Ranking states on student achievement outcomes.*

| | Improving Black student achievement | | Closing the achievement gap | |
|---|---|---|---|---|
| Ranking | Lower poverty stratum (S1) | Higher poverty stratum (S2) | Lower poverty stratum (S1) | Higher poverty stratum (S2) |
| 1 | KY, MD, NY, NC | NY, NC | NY | CA, NY |
| 2 | SC, TX, VA | CA, KY, MD, MI, SC | CA, MD | KY, NC, SC |
| 3 | CA, MI, TN | TN, TX, VA | KY, MI, NC, SC, TN, TX, VA | MD, MI, TN, TX, VA |

We observe that the rankings of two states are consistent across outcomes: New York is in the highest category in all four rankings, while Tennessee is in the lowest category in all four rankings. Most states, however, present a somewhat mixed picture. North Carolina fares well in both strata with respect to improving Black student achievement but poorly in closing the achievement gap. California fares better with respect to both outcomes in S2 than in S1. Michigan, Texas and Virginia appear in the lowest category in three of the rankings. South Carolina is assigned the middle category in three of the rankings. Finally, Kentucky and Maryland have the most heterogeneous assigned categories.

Next, we examine patterns in the rankings for the two strata within each outcome. With respect to "Improving Black student achievement", we observe a modest consistency across strata: New York and North Carolina are in the highest category in both strata, Tennessee is in the lowest category in both strata, and the other seven states differ by only one level between strata. For the outcome "Closing the Achievement Gap," there is again a modest consistency across strata: New York is in the highest category in both strata; Michigan, Tennessee and Virginia are in the lowest category in both strata; and the other six states differ only by one level between strata.

## Results of HLM Analyses

In this section we will discuss the results of carrying out a series of analyses based on HLMs. The analyses were carried out for both strata in all ten states for the years 1992 and 2000. This section presents estimates of the achievement gap employing different models, as well as a number of comparisons of trends in achievement gaps. These more complex models offer some insights into the structure of the achievement gap. Unambiguous interpretations are elusive because of the complexity of the dynamics among the different factors. Moreover, patterns of school segregation combined with limitations of the NAEP sample constrain the generalizability of the results for some states.

We have already indicated some of the caveats that must be kept in mind when interpreting the results of the within-stratum comparisons over time between Black students and White students.

In particular, it has been well documented that there can be substantial differences in average achievement among schools and that students with similar characteristics attending different schools achieve at different levels. Such school/peer effects (to the extent they exist) are confounded with our estimates of the Black-White gaps within strata, because Black students and White students have differential patterns of enrollment across schools within a stratum.

It is natural, then, to ask what the gap estimates would be if it were possible to eliminate the contribution of between-school differences. We turn for answers to the method of hierarchical linear models (HLM). An HLM can include a contrast at the student level of the model that enables us to estimate the average difference in achievement between Black students and White students attending the same schools in a particular stratum. This so-called "pooled within-school" estimate of the Black-White gap is free of average between-school differences.

It is important to recognize that two stages of data restriction are involved. In the first stage, we exclude schools whose NAEP samples have neither White students nor Black students. The second stage occurs within the HLM analysis. Specifically, only those schools with NAEP samples that included both Black and White students can contribute to the estimation of the contrast of interest. For some combinations of stratum/state/year the numbers of schools and students contributing to the estimate were quite small. This is especially the case for S2 in 1992.

Thus, while the prospect of obtaining estimates of achievement gaps free of between-school differences is an exciting one, it comes at considerable cost: The sample of schools on which such estimates are based can be relatively small and the sample is not generally representative of the stratum as a whole. Moreover, comparisons of these estimated gaps among states is more problematic because of the different degrees of *de facto* school segregation (within a given stratum) from state to state. Nonetheless, we believe these estimates should be of interest to state policy makers, as they do present a picture of within-school differences in achievement for that subset of the schools that enrolled both Black and White students.

## *Black-White Achievement Gaps*

Table 16 presents new counts of schools, White students, and Black students for each stratum/state combination, for 1992 and 2000. These are schools for which the NAEP sample contained both Black and White students.[18] Only the students in these schools contribute to the estimate of the within-school achievement gap.

[18] Presumably, there were schools participating in NAEP that had both White students and Black students, but for which the NAEP student sample did not happen to include students of one or the other race. Such schools do not appear in Table 16.

Table 16
*Number of schools and students by state (Reduced school sample).*

| | | | 1992 | | | | 2000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | # of Students | | | | # of Students | | |
| State | | # of Schools | Total | White | Black | # of Schools | Total | White | Black |
| CA | S1 | 38 | 954 | 451 | 84 | 24 | 544 | 255 | 44 |
| | S2 | 7 | 250 | 49 | 36 | 16 | 434 | 66 | 56 |
| KY | S1 | 50 | 1348 | 1093 | 190 | 38 | 913 | 747 | 129 |
| | S2 | 8 | 199 | 137 | 55 | 23 | 549 | 406 | 118 |
| MD | S1 | 70 | 1839 | 1119 | 515 | 68 | 1606 | 1019 | 382 |
| | S2 | 5 | 111 | 21 | 81 | 11 | 219 | 34 | 158 |
| MI | S1 | 31 | 826 | 653 | 113 | 22 | 550 | 455 | 32 |
| | S2 | 9 | 215 | 60 | 118 | 9 | 167 | 57 | 77 |
| NY | S1 | 27 | 705 | 466 | 127 | 22 | 506 | 349 | 53 |
| | S2 | 7 | 177 | 54 | 50 | 19 | 390 | 83 | 117 |
| NC | S1 | 74 | 1989 | 1395 | 498 | 64 | 1484 | 1002 | 373 |
| | S2 | 16 | 435 | 157 | 239 | 21 | 469 | 146 | 260 |
| SC | S1 | 73 | 1885 | 1271 | 491 | 46 | 1127 | 804 | 257 |
| | S2 | 23 | 586 | 213 | 321 | 41 | 981 | 449 | 467 |
| TN | S1 | 50 | 1354 | 1054 | 258 | 48 | 1211 | 914 | 218 |
| | S2 | 7 | 184 | 87 | 90 | 13 | 285 | 138 | 128 |
| TX | S1 | 50 | 1292 | 720 | 173 | 42 | 1003 | 554 | 150 |
| | S2 | 15 | 482 | 126 | 67 | 18 | 509 | 96 | 102 |
| VA | S1 | 85 | 2251 | 1552 | 483 | 78 | 1868 | 1255 | 403 |
| | S2 | 7 | 172 | 89 | 62 | 13 | 258 | 84 | 134 |

Comparing Table 16 to Table 9, we observe that in both years there is a substantial reduction in numbers for all states, with the exception of North Carolina, South Carolina and Virginia. The reduction in the S2 school sample in 1992 resulted in only three states (North Carolina, South Carolina, and Texas) having ten or more schools contributing to the estimation of the contrast. Consequently, the gap estimates for S2 in 1992, based on these reduced samples, have greater uncertainty attached to them. In 2000, however, only one state (Michigan) had fewer than ten schools contributing to the estimation of the contrast.

As explained in the methods section, we fit two types of HLMs, yielding what we termed adjusted and fully-adjusted estimates of the achievement gap[19]. In Table 17 we display four different estimates of the Black-White gaps for S1 in 1992 and 2000. For 1992, column (1) contains the estimated gaps already presented in Table 14, which compare weighted estimates of the mean achievement of all White students and of all Black students in the stratum. (The relevant counts are found in Table 9.) The estimates in the next three columns are based on the reduced school sample, which is delineated in Table 16. They all employ unweighted analyses. Column (2) contains the differences in the average scores of White students and Black students. Column (3) displays the adjusted achievement gaps from the first type of HLM, while column (4) displays the fully-adjusted

[19] The first type adjusts student scores for differences in schools attended; the second adjusts for schools attended, as well as student SES and student academic focus.

achievement gaps from the second type of HLM. This pattern is repeated for 2000 in columns (5) through (8). Table 18, for the data from S2, has a parallel structure to Table 17. (Note that the ordering of the states is the same as that in Table 14)

It is important to recognize at the outset that the estimates based on the full school sample are not directly comparable to estimates based on the reduced school sample both because of differences in the samples and because of the use of weights in the former but not the latter.[20] Accordingly, in what follows we will focus on comparisons among estimates based on the reduced school sample.

We first observe that in S1, in both 1992 and 2000, removing school effects results in a reduction in the estimated gaps. For example, in New York the reductions are 30 –21 = 9 points and 26 –17 = 9 points in 1992 and 2000, respectively. In some states the reductions are substantial. In 1992, five states had reductions of at least nine points; in 2000 there were three such states. Turning to S2, removing school effects also results in reductions in the estimated gaps. (The only exceptions are Kentucky and Maryland in 2000.) Again, some states exhibit very substantial reductions in both years. In 1992, three states had reductions of at least ten points; in 2000 there were four such states.

It is noteworthy that even after removing school effects the estimated gaps in each stratum/state/year combination are quite large, with most exceeding 20 points. That is, the typical pooled within-school achievement gap is about the same size as the typical gap between strata (see Table 11). That these within-school gaps are both relatively large and persistent, points to the need to probe more deeply into the differences in personal characteristics and school experiences that may account (in a statistical sense) for some portion of the achievement gap within schools.

When we adjust the estimate of the pooled within school achievement gap for differences among students in SES and AcadFoc, the gap is decreased in 1992 (column 4) for both S1 and S2, but only very slightly in 2000 (column 8). This may be due to the fact that differences in AcadFoc between White students and Black students were reduced from 1992 to 2000, so that variable was unable to account for much of the differences in achievement between the two groups.

[20] Interestingly, for S1, the estimates in columns (1) and (2) are quite similar for most states in both 1992 and 2000. The medians across states are nearly identical. This is essentially the case for S2 in 1992 as well, but not in 2000.

Table 17

*Lower poverty stratum: Black-White achievement gaps from four analyses.*

| | 1992 | | | | 2000 | | | |
|---|---|---|---|---|---|---|---|---|
| | Full school sample | Reduced school sample | | | Full school sample | Reduced school sample | | |
| | Include school effects | Include school effects | Remove school effects | Remove school effects and student covariates | Include School effects | Include school effects | Remove school effects | Remove school effects and student/ school covariates |
| State | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| NY | 29 (4.9) | 30 (1.6) | 21 (3.6) | 14 (3.2) | 15 (7.1) | 26 (2.5) | 17 (5.2) | 15 (4.7) |
| CA | 37 (3.6) | 32 (2.3) | 15 (2.2) | 10 (2.8) | 32 (5.7) | 32 (1.5) | 15 (3.9) | 11 (4.5) |
| SC | 30 (1.6) | 29 (0.6) | 26 (1.6) | 17 (1.4) | 30 (3.1) | 30 (1.1) | 26 (2.8) | 22 (2.6) |
| KY | 24 (3.3) | 25 (0.5) | 23 (3.5) | 16 (2.5) | 23 (3.4) | 24 (1.4) | 24 (3.5) | 23 (3.1) |
| MI | 37 (3.4) | 29 (0.9) | 23 (4.0) | 21 (3.2) | 38 (5.7) | 37 (1.8) | 33 (6.2) | 30 (5.4) |
| TX | 34 (2.8) | 34 (1.3) | 22 (2.4) | 14 (2.1) | 34 (4.5) | 31 (0.6) | 18 (2.2) | 12 (2.4) |
| MD | 35 (2.4) | 35 (0.6) | 26 (1.8) | 18 (1.8) | 33 (2.5) | 31 (1.3) | 25 (2.2) | 23 (2.1) |
| VA | 29 (2.0) | 31 (1.0) | 22 (2.1) | 15 (2.0) | 31 (2.5) | 29 (0.8) | 21 (2.0) | 17 (1.8) |
| TN | 27 (2.2) | 29 (0.6) | 28 (2.0) | 21 (1.6) | 28 (6.1) | 28 (1.4) | 24 (3.1) | 20 (2.9) |
| NC | 25 (1.8) | 27 (0.8) | 26 (1.7) | 19 (1.4) | 34 (2.1) | 33 (0.7) | 29 (2.4) | 22 (2.1) |
| Median | 29 | 30 | 23 | 16 | 31 | 31 | 24 | 22 |

Table 18

*Higher poverty stratum: Black-White achievement gaps from four analyses.*

| | 1992 | | | | 2000 | | | |
|---|---|---|---|---|---|---|---|---|
| | Full school sample | Reduced school sample | | | Full school sample | Reduced school sample | | |
| | Include school effects | Include school effects | Remove school effects | Remove school effects and student covariates | Include school effects | Include school effects | Remove school effects | Remove school effects and student/school covariates |
| State | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| NY | 43 (6.9) | 34 (2.5) | 18 (4.5) | 14 (4.1) | 31 (4.5) | 25 (0.9) | 7 (3.1) | 5 (3.6) |
| CA | 43 (2.4) | 34 (2.7) | 13 (3.5) | 8 (4.6) | 30 (6.7) | 23 (2.6) | 3 (2.8) | 8 (3.4) |
| SC | 27 (2.8) | 26 (1.0) | 24 (2.9) | 18 (2.2) | 24 (2.6) | 24 (1.2) | 22 (2.3) | 20 (2.1) |
| KY | 21 (5.3) | 24 (2.6) | 23 (4.1) | 15 (5.2) | 16 (4.6) | 18 (0.9) | 20 (4.3) | 19 (3.6) |
| MI | 39 (5.7) | 30 (2.0) | 16 (7.5) | 16 (6.7) | 39 (4.8) | 25 (1.6) | 12 (5.4) | 16 (5.6) |
| TX | 33 (4.5) | 30 (2.5) | 15 (4.0) | 8 (5.1) | 34 (3.8) | 32 (1.9) | 14 (3.7) | 18 (3.3) |
| MD | 26 (11.5) | 23 (2.5) | 17 (8.7) | 12 (5.4) | 24 (10.6) | 17 (3.3) | 22 (7.9) | 23 (6.5) |
| VA | 20 (6.7) | 21 (2.4) | 20 (3.9) | 12 (3.4) | 32 (6.1) | 26 (1.5) | 19 (3.9) | 15 (3.8) |
| TN | 29 (3.6) | 25 (2.1) | 22 (6.6) | 17 (4.1) | 31 (4.5) | 30 (2.9) | 22 (8.8) | 21 (6.9) |
| NC | 31 (3.7) | 30 (0.7) | 24 (3.9) | 16 (3.4) | 28 (3.5) | 24 (1.4) | 20 (3.0) | 17 (3.0) |
| Median | 30 | 28 | 19 | 14 | 31 | 24 | 20 | 19 |

Table 19

*Differences in achievement gaps between strata [S1 – S2], within state, by year (full and reduced school samples).*

| | 1992 | | | 2000 | | |
|---|---|---|---|---|---|---|
| | Full school sample | Reduced school sample | | Full school sample | Reduced school sample | |
| State | Include school effects | Remove School effects | Remove school effects & student covariates | Include school effects | Remove school effects | Remove school effects & student covariates |
| CA | -6 | 2 | 2 | 2 | 12 | 3 |
| KY | 3 | 0 | 2 | 7 | 4 | 4 |
| MD | 9 | 9 | 6 | 9 | 3 | 0 |
| MI | -2 | 7 | 5 | -1 | 21 | 14 |
| NC | -6 | 2 | 3 | 6 | 9 | 5 |
| NY | -14 | 3 | 0 | -16 | 10 | 10 |
| SC | 3 | 2 | -1 | 6 | 4 | 2 |
| TN | -2 | 6 | 4 | -3 | 2 | -1 |
| TX | 1 | 7 | 6 | 0 | 4 | -6 |
| VA | 9 | 2 | 3 | -1 | 2 | 2 |
| Median | -0.5 | 2.5 | 3 | 1 | 4.5 | 2.5 |

Table 19 presents, for each year, the differences in achievement gaps between strata, within a state.[21] Results are displayed for the descriptive analysis, as well as both HLM-based analyses. It is evident that, with school effects removed, the achievement gap in the lower poverty stratum (S1) is uniformly higher than in the higher poverty stratum, although no individual state difference is statistically significant. There is a suggestion of the differences between strata being larger in 2000 than in 1992, with substantial increases in California and Michigan. In California, this is a consequence of a reduction in the gap in S2 from 1992 to 2000, while in Michigan it is a consequence of an increase in the gap in S1 coupled with a small reduction in S2. The patterns in between stratum differences and trends over time are similar for the results from the analysis in which both school effects and student covariates are removed.

We now turn to trends over time in the achievement gap within strata. Results are presented in Table 20. The entries represent the reduction in the gap between 1992 and 2000, with positive numbers signaling a reduction and negative numbers an increase.[22] Focusing first on the results from removing school effects only, we note that for S1 there is a nearly even balance between reductions and increases in the gap, with a median of 0.5 points. This is similar, overall, to the results presented in Table 14, although there are discrepancies for particular states. For example, in the full sample, New York experienced a reduction of 14 points, while for the HLM analysis (based on the reduced school sample) the gap decreased by only 4 points. For S2, eight out of the ten states experienced reductions, with a median reduction of about 2.5 points. Again, this is similar to the results presented in Table 14, where the median reduction was about 2 points. With the reduced school

[21] That is, the tabled entry = [White mean –Black mean]$_{S1}$ –[White mean – Black mean]$_{S2.}$

[22] That is, the tabled entry = [White mean – Black mean]$_{1992}$ – [White mean – Black mean]$_{2000}$.

sample, both California and New York experienced the greatest reductions in the gap—as was the case with the full school sample.

Table 20
*Reduction in achievement gap within stratum by state [1992 to 2000]. (Reduced school sample.)*

| | S1 | | S2 | |
|---|---|---|---|---|
| State | Remove school effects | Remove school effects & student covariates | Remove school effects | Remove school effects & student covariates |
| NY | 4 | -1 | 11 | 9 |
| CA | 0 | -1 | 10 | 0 |
| SC | 0 | -5 | 2 | -2 |
| KY | -1 | -7 | 3 | -4 |
| MI | -10 | -9 | 4 | 0 |
| TX | 3 | 2 | 1 | -10 |
| MD | 1 | -5 | -5 | -11 |
| VA | 1 | -2 | 1 | -7 |
| TN | 4 | 1 | 0 | -4 |
| NC | -3 | -3 | 4 | -1 |
| Median | 0.5 | -2.5 | 2.5 | -2.5 |

When we consider the results for the analyses in which the estimated gaps are adjusted for student covariates, as well as for school effects, we observe a preponderance of increases in both gaps for strata. The median increase in the gap is about 2.5 points. Again, there is some ambiguity in interpreting these results, as they are influenced by the differential trajectories of the covariates over time.

## Linking Results

In an earlier section we presented rankings of states into one of three categories on each of five policy levers, as well as an overall ranking (Table 7). Subsequently, we presented the results of our analyses of NAEP score trends drawing on data from the full school samples. Table 15 displays the ranking of states based on those results.

Although policy makers are invariably interested in "what works", a statistical analysis of the available data cannot answer this question directly. Patterns of association between policy rankings and outcome rankings can, however, provide useful insights. Specifically, it is possible to determine whether those states judged to be more successful in the policy realm were among those that experienced greater improvements in student outcomes and, conversely, whether those judged to be less successful tended to experience poorer outcomes. Given the nature of the phenomena under study, neither one policy analysis nor a single set of statistics can adequately convey the many strands that constitute a state's "story" or the complexity of the relationships among the strands for ten different states. We have chosen, therefore, to adopt multiple perspectives.

In this context, an analysis of the possible impact of educational policy requires consideration of the trajectories of both groups, as well as of the achievement gap between them. Accordingly, we employ two outcome measures, improvements in the NAEP scores of Black students and ieductions in the achievement gaps between Black students and White students. For each measure, we examine the state's record in both the higher poverty and lower poverty strata. In our study of the linkage between the outcome measures and the states' education policies, we consider the overall policy ranking, as well as the rankings on each policy component. Recall that even though the rankings are grounded in the extensive data collected and analyzed, there is necessarily some subjectivity involved in the process.

To assist in the evaluation of the strength of the linkage between policy and outcomes, we rely on a simple measure of association: We count the number of category shifts that would be necessary to bring a policy ranking and an outcome ranking into perfect alignment—the fewer the number of shifts, the stronger the linkage.[23]

## *The Search for Patterns*

Consider first the outcome "Improving Black student achievement." The "Overall" ranking, as well as the rankings on "Teacher quality" and "Assessment and accountability" provide the best matches to the combined (across the two strata) rankings on this outcome. The matches for the other three policy components are clearly poorer. The pattern of association between the "Overall" ranking and the outcomes for the two strata is represented in Figure 6. Note that, ideally, states' rankings on both dimensions would be identical. In that case, off-diagonal cells would be empty.

| Overall Policy Rankings | Improving Black Student Achievement—Rankings | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | |
| 1 | nc | NC | | | | |
| 2 | ky<br>ny | NY | sc<br>tx | CA<br>KY<br>SC | ca | TX |
| 3 | md | | va | MD<br>MI | mi<br>tn | TN<br>VA |

*Figure 6*. Relationship between overall policy ranking and improving Black student achievement. Cell entries are states: lower case abbreviation denotes the lower poverty stratum and upper case abbreviation denotes the higher poverty stratum.

One can argue that the quality of the match is nearly as good as is possible, given the pairs of outcome ranks associated with each state. For the "Overall" ranking, the state in the highest category is also rank 1 on the outcome in both strata, four of the five states in the middle category

[23] This count is approximately a non-normalized version of Kendall's tau, a measure of association often used with ranked data.

were also rank 2 on the outcome in at least one stratum, and three of the four states in the lowest category were also rank 3 in at least one stratum and never in rank 1. Maryland and New York have the greatest disparities between their overall policy ranking and their ranking on outcomes. Maryland is in the lowest policy category but is in rank 1 and rank 2 with respect to the outcome. New York is solidly in the middle category on policy, but very clearly rank 1 on the outcome in both strata.

With respect to the outcome "Closing the Achievement Gap", the "Overall" ranking provides the best match to the combined rankings. Among the five policy levers, the rankings on "Curriculum and standards" and "Teacher quality" have the best matches, while "Governance" and "Assessment and accountability" have the poorest matches. The pattern of association between the "Overall" ranking and the outcomes for the two strata is represented in Figure 7. The linkage with policy is weaker for this outcome. The pattern is clearest for those states in the lowest category for overall policy—they are most likely to fall in the lowest category on the outcome for both strata.

| | Closing the achievement gap—Rankings | | | | | |
|---|---|---|---|---|---|---|
| Overall Policy Rankings | 1 | | 2 | | 3 | |
| 1 | | | | NC | nc | |
| 2 | ny | CA<br>NY | ca | KY<br>SC | ky<br>sc<br>tx | TX |
| 3 | | | md | | mi<br>tn<br>va | MD<br>MI<br>TN<br>VA |

*Figure 7*. Relationship between overall policy ranking and improving Black student achievement. Cell entries are states: lower case abbreviation denotes lower poverty stratum and upper case abbreviation denotes higher poverty stratum.

As the evaluation of states' policies makes clear, each state focused on a different set of policy levers—with different degrees of success—over the period of interest. The patterns just described indicate that there is no simple accounting for states' results. No single dimension predicted states' outcomes as well as the overall ranking did, although *teacher quality* was nearly as good. Policy rankings are more predictive of states' ranks on improving Black achievement than of their ranks on closing the achievement gap. Not only is the latter outcome more refractory to progress but also it is less obvious what combination of policies can lead to some success.

## *Reducing the Gap: Selected State Stories*

Turning to individual states, New York, with the best record on outcomes, was rated in the highest category on teacher quality, and in the middle category on the other four policy components, as well as on overall policy. During this period, New York was distinguished by its greater emphasis on professional development for its mathematics teachers than on high-stakes accountability. (It serves as a counter-example to the claim that high stakes testing is essential to making progress.) New York did have a solid assessment program that had been under development from the mid-1980s, as well as a rigorous end-of-high school examination battery (Regents' Examinations) that was not then required for graduation during 1988–98. Nonetheless, it was thought to exert some influence throughout the system.

California was second to New York in reducing the achievement gap. Overall, it was more successful in the higher poverty stratum than in the lower poverty stratum. Like New York, it was rated in the middle category on four of the five policy components, and in the highest category on education finance. California did not institute high-stakes testing and, in fact, its assessment program suffered considerable discontinuities during the period.

North Carolina was distinguished by its placement in the highest category on four policy components and the fact that it substantially raised Black student achievement in both strata. (However, some of this improvement should be discounted in view of the state's large increase in exclusion rates.) With respect to closing the achievement gap, on the other hand, North Carolina's record is rather poor.

Kentucky's policies have been well documented and, in many ways, it was considered a leader in education reform in the 1990s. Although it began with a comprehensive design, its implemented reforms were incomplete and beset by political battles. It had limited success with respect to either outcome.

Texas, which employed high stakes testing as the driving force behind its education reforms, had relatively little success during the period of interest. Contrasting the data from Texas, on the one hand, with that of New York and California on the other, suggests that test-based accountability alone cannot carry the day. Rather, a reasonably consistent, broad-based effort can be more successful.

Maryland was notable for its consistent and somewhat narrow focus on its state assessment program, the Maryland School Performance Assessment Program (MSPAP), Initiated in 1991, the MSPAP was the state's primary reform during this period. The state did not expend a great deal of energy on improving fiscal equity between schools and districts, was relatively slow in developing extensive standards for teachers to refer to, and did not do a great deal with respect to raising teacher quality or qualifications. However, because the state is small (with only 24 school districts), the state department was able to maintain good communication channels with every school district.

Compared to the other states in the study, Virginia was late in introducing quality state standards and an accountability system. It was only in 1994 that Virginia began to overhaul the state system of standards and accountability. Starting in 1995, the state undertook a significant effort to familiarize all teachers with what is known as the Standards of Learning (SOLs), a set of standards and objectives for each grade and course. The state funded a lead teacher in each school in order to train teachers and oversee implementation of the SOLs. Assessments aligned with the SOLs were introduced in 1998, toward the end of our period of study.

South Carolina had a different strategy than all the other states in this study. It began in 1993 with a slow introduction of standards and programs to help teachers become familiar with the new standards and practices, followed by progressive upgrading of teacher quality standards and some attempts at equalizing funding for poorest school districts. Then, in 1998, the state introduced aligned assessments and increased accountability. The ongoing changes in terms of standards and improvement programs proved to be confusing to some teachers and administrators who felt they were being asked to make incremental changes over many years.

The strong tradition of local control in Michigan, combined with the lack of explicit support from both the teachers' union and the higher education institutions in the state, meant that changes introduced by the state education department had limited impact on local districts. The two main reforms in the state were the introduction of the core content standards in 1990 and the 1993–1994 overhaul of school finance.

Tennessee, in spite of its finance reform initiated in 1993, remained the lowest spending state in terms of per-pupil expenditure of any of the states in this study, and remains well below the national average. Combined with standards that varied in quality during the time period in question, no accountability, and an unclear link between standards and assessment, it is not surprising that the state experienced relatively poor outcomes over the period.

## Discussion

The challenge for American education is both to raise achievement for all and to close the gaps between majority and minority students. As various education reforms are developed and launched (often with considerable fanfare), the paramount question is: Has there been any improvement and, if so, to what can we attribute the success? The first part of the question is amenable to statistical analysis. The second part, concerning cause and effect, cannot be answered definitively. The difficulty is both that the data are derived from observational surveys and that the different policies overlap in time. Consequently, it is almost impossible to isolate the impact of a particular policy.

This study describes an approach that is at once limited and ambitious. It considers the experiences of just ten states over a decade and focuses on analyzing the NAEP scores of White students and Black students during that period on a single assessment in one grade. On the other hand, the public schools in the ten states in this study enrolled about 40–45% of all Black students in public schools in the nation. The study breaks "new ground" by probing beneath the surface of reported state results to estimate the achievement gaps within poverty strata and even within schools within strata. Further, it incorporates extensive policy histories for each state, looking for links between states' policies and their relative success in meeting the challenges cited above.

With respect to outcomes, mean achievement rose in both poverty strata in all ten states—for all students, for White students and for Black students. There was, however, considerable variation across states in the amount of improvement. Further analysis revealed substantial heterogeneity within states: The typical achievement gap between strata within states was about 20 points and the typical Black-White achievement gap within a stratum was about 30 points. For nearly all states, the within-stratum Black-White achievement gaps were very similar in magnitude for the two strata. What was more surprising was that for most states the within-stratum Black-White achievement gaps were nearly as large as the Black-White achievement gap for the state as a whole. That is, the large differences in average scores between lower poverty and higher poverty schools accounted for only a part of the state level Black-White achievement gap. This finding is of some interest in view of suggestions of using social class rather than race as a basis for affirmative action.

To put these results in perspective, in 1992 the difference between the highest state mean and the lowest was only ten points. Yet across the ten states, the median Black-White achievement gap within each stratum of about 30 points is nearly as large as the standard deviation of test scores in the full population, corresponding to an effect size of almost one. Finally, in 1990, when the cross-grade scale in mathematics was established, the difference in scores between the median 8th grader and the median 4th grader was 50 points.

The focus of the study has been on the trajectories of the achievement gaps from 1992 to 2000. Across the ten states, the median gap between strata and the median Black-White achievement gap within a stratum each remained essentially constant over the period. In fact, for most states there was little change in the outcomes we considered. On the other hand, there were considerable differences in trajectories for some states, which led to the question of whether differences in states' education policies could account for the variation in states' results.

To address this question, we developed a policy framework and compiled extensive histories of states' policies from (roughly) 1988 to 1998, supplemented by information provided by experts in each state. We found that states' overall policy rankings correlated moderately with their record in improving Black student achievement but were less useful in predicting their record in reducing the Black-White achievement gap. For both outcomes, the strength of the association between policy and the direction of the trajectory was limited by the fact that, for most states, results differed by stratum.

No single policy component accounted for the differences among states in the two outcomes as well as did the overall policy rating—although *teacher quality* did nearly as well. Perhaps the clearest finding was that states in the lowest category on overall policy (Michigan, Tennessee and Virginia) were also in the lowest category in both strata with respect to reducing the achievement gap.

One caution in interpreting the results on the Black-White achievement gaps within strata is due to the differential distributions of Black students and White students across schools within a stratum. Thus, observed gaps could be confounded with average differences among schools. Accordingly, we undertook a series of multi-level analyses to estimate the size of the achievement gap, after eliminating between-school differences. The resulting pooled, within-school estimates were somewhat smaller than the original estimates, but still quite large. In this setting, we observed some interesting consistencies. For all states in both years, the achievement gap in S1 was greater than the achievement gap in S2. For most states, the achievement gap was reduced in both strata—particularly in S2.

## *Limitations*

There are a number of limitations that should be taken into account in interpreting these results. First, in carrying out the policy analysis under time and budget constraints, we could not capture all the salient features of the political environment and policy landscape in the ten states over a ten-year period. To do so would have required writing a book on each state—as has been done for California and Kentucky! Such a book would have included interviews with a larger number of experts, more information on the scope and quality of the implementation of the various reform initiatives, as well as data on demographic and economic trends within the state. Our categorizations of the states on the different policy levers were necessarily grounded in the information that we were able to amass and, ultimately, determined by our subjective judgments of the relative strengths of each state's efforts.

With respect to the data analysis, our estimates of the trajectories are potentially confounded with differences between the 1992 and 2000 cohorts. It is also the case that NAEP was not designed to support the kind of "deep" analyses we have carried out. The target number of schools in a state sample was only about 100 and, in some cases, the realized sample was noticeably smaller. Furthermore, the number of students tested within each school is only about 20. Consequently, the sample sizes for estimating the achievement gaps in many stratum/state/year combinations were disappointingly small and the estimated standard errors correspondingly large.

Because the analyses were exploratory and intended to lead to a delineation of summary patterns, we used the standard errors for guidance, rather than hewing to the conventional .05 level for significance. Nonetheless, it should be borne in mind that the categorizations of the states were based on estimated changes over time that typically were not significantly different from zero and were subject to considerable uncertainty.

Small sample sizes in the reduced school sample were particularly problematic when using HLM to estimate the pooled within-school achievement gap. Moreover, because of the way in which schools were originally selected for NAEP, the estimates derived from the reduced school sample cannot be directly generalized to a larger population of schools. Thus, those results must be treated as suggestive of what one might have found in a school sample expressly drawn to estimate the pooled within-school achievement gap.

This study employed data from the NAEP 8th grade mathematics assessment for ten states. We are not in a position to generalize our findings to other states, other subjects or other grades. Finally, we must keep in mind that state policies are focused on strengthening the state's curriculum frameworks and standards, enhancing the alignment of instruction with those standards, and improving scores on the state's tests. To the extent that the NAEP mathematics framework and the NAEP assessment differ from the state's, the patterns in NAEP results may not be an entirely fair basis for making judgments about the relative success of the state's education reform efforts. Furthermore, the magnitude of the achievement gap and its trajectory over time will vary with the choice of the criterion. For example, trends in the difference in the proportions of the two groups meeting the state's standard of proficiency can be quite unlike trends in the difference in the means for the two groups—whether on a state test or on a comparable NAEP test.

## *Conclusions*

We believe that this study supports the assertion that analyses of outcomes within states are more interesting and useful than gross comparisons among states. From 1992 to 2000, Black students in the eighth grade made modest gains on the NAEP mathematics assessment but, overall, the Black-White achievement gap remained essentially constant—at each of the three levels of analysis we carried out. We conclude that the Black-White achievement gap can be fairly characterized as pervasive, profound and persistent: Pervasive because the gap exists in all ten states, profound because the gap can be found at all three levels of aggregation, and persistent because overall the gap did not diminish over the period of study.

While a few states made some progress and others lost ground, in all our analyses most states experienced little change in the achievement gap over the period of the study. Nonetheless, our categorizations of the states yielded an interesting result: Policy variations among states appear to account for a modest amount of the differences in outcomes. In other words, top-down reform is a blunt tool, but a tool nonetheless. At the same time, these findings highlight the difficulties in realizing some of the goals that inspire state reforms. Although individual schools may demonstrate progress, large-scale effective reforms appear to be much rarer. While there are many reasons why this is the case, a particularly constructive analysis of the difficulties associated relying on standards-based reform to effect school improvement can be found in O'Day (2002).

On the basis of this study, we cannot make facile pronouncements of "what works." Our findings are consistent with the recommendations made by the Center on Education Policy (2001) for closing the achievement gap. We want to draw particular attention to its warning, "Closing the gap will require bold, comprehensive and long-term strategies" (p. iii). In view of our within-school results, "bold" must include policies that directly support local reform efforts with demonstrated effectiveness in addressing the experiences of students of different races attending the same schools. Considering our overall findings, "comprehensive" should signify full use of all policy levers, rather than reliance on one or two. Coherence and consistency do matter. Finally, "long-term" means that reform efforts should be built on broad-based support and structured to be better able to withstand the vicissitudes of economic trends and state politics.

## References

Allen, N. L., Jenkins, F., Kulick, E., & Zelenak, C. A. (1997). *Technical report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: National Center for Education Statistics.

Allen, N. L., Johnson, E. G., Mislevy, R. J., & Thomas, N. (1999, July). Scaling procedures. In N. L. Allen, J. E. Carlson & C. A. Zelenak (Eds.), *The NAEP 1996 technical report, NCES 1999–452* (pp. 75–95). Washington, DC: National Center for Education Statistics.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved March 28, 2002, from http://epaa.asa.edu/epaa/v10n18/.

Antonucci, M. (1999). *Measure for measure: A magnified look at standardized test scores.* Carmichael, CA: Education Intelligence Agency.

Barton, P. E. (2002). *Raising achievement and reducing gaps: Reporting progress toward goals for academic achievement in mathematics.* Washington, DC: National Education Goals Panel.

Barton, P. E. (2003, October). *Parsing the achievement gap.* (Policy Information Report). Princeton, NJ: Educational Testing Service.

Barton, P. E., & Coley, R. J. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade.* (Policy Information Report). Princeton, NJ: Educational Testing Service.

Blank, R. K. (2000). *Summary of findings from SSI and recommendations fro NSF's role with states: How NSF can encourage state leadership in improvement of science and mathematics education.* Washington, DC: Council of Chief State School Officers.

Bracey, G. W. (2002). International comparisons: An excuse to avoid meaningful educational reform. *Education Week, 21*(19), p. 30.

Bracey, G. W. (2003). *Those misleading SAT and NAEP trends: Simpson's paradox at work.* Retrieved September 21, 2004, from http://www.americatomorrow.com/bracey/EDDA/EDDRA30.htm

Brady, R. C. (2003). *Can failing schools be fixed?* Washington, DC: Thomas B. Fordham Foundation.

Braun, H. I. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*(1). Retrieved September 22, 2004, from http://epaa.asu.edu/epaa/v12n1/.

Burke, J., & James, P. (1997). Weighting procedures and variance estimation. In N. L. Allen & et al. (Eds.), *Technical report of the NAEP 1996 state assessment program in mathematics*. Washington, DC: National Center for Education Statistics.

Camilli, G. (2000). Texas gains on NAEP: Points of light? *Education Policy Analysis Archives, 8*(42). Retrieved September 22, 2004, from http://epaa.asu.edu/epaa/v8n42/.

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–331.

Cawelti, G. (1999). *Portraits of six benchmark schools: Diverse approaches to improving student achievement.* Arlington, VA: Education Research Service.

Center on Education Policy. (2001). *It takes more than testing: Closing the achievement gap*. Washington, DC: Center on Education Policy.

Chambers, R. L. (2003). *Which sample survey strategy? A review of three different approaches.* (S$^3$RI Methodology Working Paper No. M03/20). Southampton, UK: Southampton Statistical Sciences Research Institute.

Christensen, L., & Karp, S. (2003, October 8). Why is school reform so hard?: Two classroom veterans rethink the process. *Education Week*.

Clune, W. H. (1998). *Toward a theory of systematic reform: The case of nine NSF Statewide Systematic Initiatives* (Research Mongraph No. 16). Madison, WI: University of Wisconsin, National Institute for Science Education.

Coleman, J. S., & et al. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Coley, R. J. (2003, November). *Growth in school revisited. Achievement gains from the fourth to the eighth grade.* (Policy Information Report). Princeton, NJ: Educational Testing Service.

College Board. (1999). *Reaching the top: A report of the National Task Force on Minority High Achievement.* New York, NY: The College Board.

Corcoran, T. B. (1997). *Evaluating systemic reform* (TEECH Paper): Teacher Enhancement Electronic Community Hall. Retrieved January 2004, from http://teech.terc.edu/papers/papers/corcoran.htm.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved March 20, 2006, from http://epaa.asu.edu/epaa/v8n1/

Dee, T. S., & Keys, B. J. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Policy analysis management, 23*(3), 471–488.

Desimone, L. M., Smith, T. M., Hayes, S. A., & Frisvold, D. (2005, Winter). Beyond accountability and average mathematics scores: Relating state education policy attributes to cognitive achievement domains. *Educational Measurement: Issues and Practices*, *24*, 5–18.

Education Trust. (1999). *Dispelling the myth: High poverty schools exceeding expectations.* Washington, DC: Education Trust.

Education Trust. (2001). *Dispelling the myth revisited. Preliminary findings from a nationwide analysis of "high-flying" schools.* Washington, DC: Education Trust.

Education Trust. (2002). *The funding gap: Low-income and minority students receive fewer dollars.* Washington, DC: Education Trust.

Edwards, V. B., & Olson, L. (Eds.) (1997, January). *Quality Counts: A report card on the condition of public education in the 50 States.* (An Education Week/Pew Charitable Trusts Report). Washington, DC: Editorial Projects in Education.

Edwards, V. B., & Olson, L. (Eds.) (1998, January). *Quality Counts '98: The urban challenge* (An Education Week/Pew Charitable Trusts Report on Education in the 50 States). Washington, DC: Editorial Projects in Education.

Edwards, V. B., & Olson, L. (Eds.) (1999, January). *Quality Counts '99: Rewarding results, punishing failure* (An Education Week/Pew Charitable Trusts Report on Education in the 50 States). Washington, DC: Editorial Projects in Education.

Edwards, V. B., & Olson, L. (Eds.) (2000, January). *Quality Counts 2000: Who should teach?* (An Education Week/Pew Charitable Trusts Report on Education in the 50 States). Washington, DC: Editorial Projects in Education.

Ferguson, R. F. (1998). Can schools narrow the Black-White test score gap? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 318–374). Washington, DC: Brookings Institute.

Friedman, B. M. (2005). Meltdown: A case study. *The Atlantic Monthly, 296*(1), 66–68.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP scores tell us.* Santa Monica, CA: RAND Corp.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved March 20, 2006, from http://epaa.asu.edu/epaa/v8n41/.

Hanushek, E. A., Rivkin, S. G., & Taylor, L. L. (1996). Aggregation and the estimated effects of school resources. *The Review of Economics and Statistics, 78*, 611–627.

Hedges, L. V., & Nowell, A. (1998). Black-White test score convergence since 1965. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 149–181). Washington, DC: Brookings Institution.

Herrnstein, R., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* Ne York: Free Press.

Hoxby, C. M. (2001). *How school choice affects the achievement of public school students.* Paper presented at the Koret Task Force meeting, Hoover Institution, Stanford, CA.

Hussar, W. & Sonnenberg, W. (2000). *Trends in disparities in school district level instructional expenditures per pupil* (NCES 2000–020). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Ingersoll, R. M. (1999). The problem of underqualified teachers in American secondary school. *Educational Researcher, 28*(2), 26–37.

Jencks, C., & Phillips, M. (1998). The Black-White test score gap: An introduction. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 1–52). Washington, DC: Brookings Institution.

Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., Robyn, A., & Burroughs, D. (2000). *Teaching practices and student achievement. Report of first-year findings from the 'Mosaic' study of systematic initiatives in mathematics and science.* (RAND Education). Santa Monica, CA: RAND.

Kober, N. (April 2001,). *It takes more than testing: Closing the achievement gap.* Washington, DC: Center on Education Policy.

Koretz, D. (2005). *Using aggregate-level linkages for estimation and validation: Comments on Thissen and Qian & Braun.* Presented at the ETS Conference on Linking and Aligning Scores and Scales: A conference in honor of Ledyard R Tucker's approach to theory and practice, Princeton, NJ, June 24–25, 2005.

Kosters, M. H., & Mast, B. D. (2003). *Closing the education achievement gap: Is Title I working?* Washington, DC: AEI Press.

Lazer, S. (1999, July). Assessment instruments. In N. L. Allen, J. E. Carlson & C. A. Zelenak (Eds.), *The NAEP 1996 technical report, NCES 1999–452* (pp. 75–95). Washington, DC: National Center for Education Statistics.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31*(1), 3–12.

Little, R. (2003, November 12). *To model or not to model? Competing modes of inference for finite population sampling.* Ann Arbor, MI: University of Michigan, School of Public Health. Retrieved January 28, 2005, from http://www.bepress.com/umichbiostat/paper4.

Ludwig, J. (2003). A review of *Educational Achievement and Black-White Inequality* by J. Jacobsen, C. Olsen, J. King Rice, S. Sweetland, and J. Ralph (Eds.). *Education Next 3*(3), 79–82. Retrieved March 11, 2005, from http://www.educationnext.org/20033/79.html.

Miller, S. (1995). *An American imperative: Accelerating minority educational advancement.* New Haven, CT: Yale University Press.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved February 20, 2006, from http://epaa.asu.edu/epaa/v14n1/.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational evaluation and policy analysis, 26*(3), 237–257.

O'Day, J. A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review, 72*(3). Retrieved February 20, 2006, from http://gseweb.harvard.edu/~hepg/oday.html.

Odden, A., & Picus, L. (2000). *School finance: A policy perspective* (2nd ed.). New York: McGraw Hill.

Pfeffermann, D., Moura, F., & Nascimento Silva, P. (2004). *Multi-level modelling under informative sampling.* ($S^3$RI Methodology Working Paper No. M04/09). Southhampton, UK: Southhampton Statistical Sciences Research Institute.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, 60*(1), 23–40.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the Black-White test score gap widen after children enter school? In C. Jenks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 229–272). Washington, DC: Brookings Institution Press.

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the Black-White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 103–146). Washington, DC: Brookings Institute.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2001). *HLM5: hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13*(2), 85–116.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Fotiu, R. P., Cheong, Y. F., & Ziazi, Z. M. (1996). Inequality of access to educational opportunity: A national report card for eighth-grade math. *Educational and Evaluation Policy Analysis, 20*(4), 256.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307–335.

Ravitch, D. (2000). *A century of failed school reforms.* New York, NY: Simon and Schuster.

Raymond, M. E., & Hanushek, E. A. (2003, Summer). High-stakes research. *Education Next, 3*(3), 48–55. Retrieved September 21, 2003, from http://www.educationnext.org/20033/.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap.* Washington, DC: Economic Policy Institute.

Smith, M. S., & O'Day, J. A. (1991). Systematic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). Bristol, PA: Falmer.

Sum, A., Kirsch, I., & Taggart, R. (2002). *The twin challenges of mediocrity and inequality: Literacy in the U. S. from an international perspective.* Princeton, NJ: Educational Testing Service.

Swanson, C. B., & Stevenson, D. L. (2002, Spring). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis, 24*(1), 1–27.

Thissen, D. (2005). *Linking assessments based on aggregate reporting: Background and issues.* Presented at the ETS Conference on Linking and Aligning Scores and Scales: A conference in honor of Ledyard R Tucker's approach to theory and practice, Princeton, NJ, June 24–25, 2005.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley Publishing Company.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform.* Cambridge, MA: Harvard University Press.

U. S. Department of Education. (1995, June). *National Education Longitudinal Study of 1988 (NELS88)--A profile of American high school seniors in 1992.* Washington, DC: National Center for Education Statistics.

U. S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics (1999). *NAEP 1999 Trends in academic progress: Three decades of student performance,* NCES No. 2000–469, by J. R. Campbell, C. M. Hombo, and J. Mazzeo. Jessup, MD: Education Publications Center (ED Pubs).

U. S. Department of Education. (2000). *Mathematics and science in the eighth grade: Findings from the Third International Mathematics and Science Study.* (NCES 2000–014.) Washington, DC: National Center for Education Statistics.

U. S. Department of Education, (2001a). *Educational achievement and Black-White inequality.* (NCES No. 2001–061.) Washington, DC: National Center for Education Statistics.

U. S. Department of Education. (2001b). *High standards for all students: A report from the national assessment of Title I on progress and challenges since the 1994 reauthorization.* Washington, DC: Planning and Evaluation Service.

U. S. Department of Education. (2001c). *Paving the way to postsecondary education: K–12 intervention programs for underrepresented youth.* (Report of the National Postsecondary Education Cooperative Working Group on Access to Postsecondary Education No. NCES 2001–205r). Jessup, MD: National Center for Education Statistics.

U. S. Department of Education. (2002). *Early Childhood Longitudinal Study--Kindergarten Class of 1998–99 (ECLS-K), Psychometric report for kindergarten through first grade.* (NCES No. 2002–05). Washington, DC: National Center for Education Statistics.

U. S. Department of Education, National Center for Education Statistics. (2004). *National Assessment of Educational Progress (NAEP), The nation's report card: Mathematics highlights 2003*. Jessup, MD: ED Pubs or http://nces.ed.gov/nationsreportcard.

Webb, N. L., Kane, J., Kaufman, D., & Yang, J.-H. (2001, June). *Study of the impact of statewide systematic initiatives program. Technical report to the National Science Foundation on the use of state NAEP data to assess the impact of the Statewide Systematic Initiatives.* Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Wilson, S. (2003). *California dreaming: Reforming mathematics education*. New Haven and London: Yale University Press.

Zucker, A. A., Shields, P. M., Adelman, N. E., Corcoran, T. B., & Goertz, M. E. (1998). *Statewide systemic initiatives program*. Washington, DC: National Science Foundation.

## About the Author

**Henry I. Braun**
Educational Testing Service

**Aubrey Wang**
School District of Philadelphia

**Frank Jenkins**
Westat, Inc.

**Elliot Weinbaum**
University of PennsylvaniaAffiliation information

Email: hbraun@ets.org

**Dr. Henry Braun**, a Distinguished Presidential Appointee at ETS, has published in the areas of probability, stochastic modeling and empirical Bayes methods. He is a co-winner of the Palmer O. Johnson award from AERA (1986) and a co-winner of the NCME award for Outstanding Technical Contributions to the Field of Educational Measurement (1991). His current interests include the interplay of technology and assessment, design science, evaluation methodology, the analysis of large-scale assessment data and education policy.

**Aubrey H. Wang** is a program evaluator at the School District of Philadelphia and the current President of the Chinese American Educational Research and Development Association. Her research interests include educational policy and practice in closing the achievement gap.

**Frank Jenkins** is a senior statistician at WESTAT, Inc, where he directs analyses of various educational and health studies involving hierarchical nesting of subjects. At ETS he worked on NAEP evaluations with particular emphasis on psychometric issues and hierarchical linear models. He has also developed Bayesian models for analyzing multivariate performance assessments.

**Elliot Weinbaum** is a researcher at the Consortium for Policy Research in Education at the University of Pennsylvania. His research interests include intergovernmental relationships and policy-making, performance-based accountability, and the impacts of policy on school improvement. He is currently an investigator in a national study of high schools and the roles that outside organizations play in high schools' strategies to improve instruction.

# Appendix A: Example of a State Profile

## Kentucky Profile

NOTE: This profile comprises six sections. Each section begins with a number of questions. Some of the questions are answered by the data in the profile, while others appear on the questionnaire we are asking you to fill out.

## *Context*

*What were the general characteristics of KY's public education system during the period 1988 through 1998?*
*How was responsibility for governing the state education system (e.g., accountability, teacher quality, curriculum and standards, finance) distributed during this period?*
*What were the most significant changes?*

### *Education System*

In 1996, Kentucky's K-12 public schools served over 650,000 students who attended school in 176 districts (CPRE, 1996). As of 2003, Kentucky has a total of 176 school districts (1,271 schools) that serve over 610,000 students (KYDE, 2003). The total number of full-time equivalent teaching staff in 2000 was 34,173 with 23,083 elementary teachers, 10,177 secondary teachers, and 913 combined elementary and secondary teachers. The average per pupil teacher ratio in 2000 was 15.64 (KYDE, 2003).

### *Governance*

*1996*. Kentucky has undergone more changes in its public education system in the last ten years than any other state in the nation (CPRE, 1996). Kentucky realized substantial achievement gains during the 1990s, after undertaking perhaps the most extensive systemic education reforms of any state in the 1990s. These included substantial equalization of school funding along with large increases in teacher salaries and overall spending; changes in school organization, including multi-age primary grade classrooms; investments in early childhood education; the introduction of standards and curriculum frameworks, along with portfolios and performance assessments. Changes in teacher education and licensing accompanied these reforms, including the adoption of the INTASC licensing standards (developed by a consortium of more than 30 states), the introduction of new licensing tests and teacher education requirements, incentives for colleges of education to meet national professional accreditation standards; and massive investments in professional development (Darling-Hammond, 2000).

As a response to the low ranking of Kentucky in 1983 in education spending per pupil, teachers' salaries, pupil-teacher ratio, high school graduation, adults with a high school diploma, and adults with a college degree (Prichard Committee, 1999; Rhoten, Carnoy, Chabran, & Elmore, 2003), in 1989, the Kentucky Supreme Court delivered a landmark decisions, ruling that the state's public school system was unconstitutional and further describing the conditions it deemed to be essential and minimal characteristics of an efficient system of common (public) schools. The General Assembly drafted the Kentucky Education Reform Act (KERA), which became law on July 13, 1990. The Act was amended in each subsequent session of the General Assembly—in 1992, 1994, and 1996 (CPRE, 1996).

The legislation addressed policy structure. The position of the elected Superintendent of Public Instruction was abolished in 1992 by an amendment to Kentucky's constitution, and almost all of this position's duties were transferred to the appointed Commissioner of Education (CPRE, 1996). KERA also abolished the existing Department of Education (referred to hereafter as the state education agency, or SEA) on June 30, 1991 and reorganized it to include new positions and a new service-oriented mission, effective July 1, 1991 (CPRE, 1996).

KERA's components include: a) educational goals indicating what graduates should know and be able to do; b) an assessment process to determine if students are reaching these goals (through the development of the Kentucky Instructional Results Information System (KIRIS); c) an accountability system holding schools responsible for student success; d) increased funding for professional development activities for educators to help students succeed; e) a new system for credentialing teachings; f) early childhood programs; g) funding to help students who require more time to achieve academic success; h) a major commitment to technology; I) full-service schools inclusive of community and agency resources; j) changes in governance structure to alter the politics in Kentucky school districts; k) as well as a commitment to fund the new initiatives (CPRE, 1996).

### *Accountability System*

After a series of conflicts over, and evaluations of, KIRIS, the process culminated in 1997 when the state fired the testing company because of scoring errors and investigated more than one hundred schools for cheating (Jacobson, 1999; Rhoten, Carnoy, Chabran, & Elmore, 2003). House Bill 53 was signed into law on April 14, 1998 and renamed the KIRIS system, the Commonwealth Accountability Testing System (CATS). The immediate political goal and its primary technical function was to revise the existing testing instruments. At the level of assessment, CATS modified KIRIS by (1) introducing the Kentucky Core Content Test, which tests students on how well they are learning the basics of math, science, reading, writing, and other subjects; (2) requiring a national, norm-referenced portion, which matches the state's core curriculum and provides national comparisons for state students; (3) calling for a pared down written portion of the test; and (4) expanding the number of grades in which these tests are administered. The state board selected CTB/McGraw Hill to run the norm-referenced component of the CTAS testing system (Rhoten, Carnoy, Chabran, & Elmore, 2003).

In addition, the new sanctions and assistance programs as prescribed by the law include: mandatory audits for struggling schools; eligibility to receive CATS school improvement money; education assistance from highly skilled, certified state staff members rather than, as under KIRIS, so-called distinguished educators, who were experienced, state-paid teachers or administrators; and the option for students at low-performing schools to transfer to successful ones (Rhoten, Carnoy, Chabran, & Elmore, 2003).

As Kentucky redesigned its system, a new education financing formula adjusted the state's district allocation from a plan that dispensed comparable funding based on student attendance and teacher experience and certification to one that varied allocations based on the amount of revenues generated by local taxes. Under this new formula, districts are required to meet certain local revenue-raising benchmarks; however, those with small tax bases and/or limited property values are protected and theoretically equalized with additional state funding. As a result, according to *Education Week*, Kentucky has accomplished a level of relative equity in spending per student, with a variation rate of only 13 percent between districts comparing to that national average of 23.1 percent. From 1989-90 to 1998-99, Kentucky reduced the gap in per pupil expenditures between wealthy and poor districts by 36.9 percent while raising the state's national rank in per pupil spending from 42 to 31 (Heine, 2002; Rhoten, Carnoy, Chabran, & Elmore, 2003).

### *Credentialing of Teachers*

Under the Kentucky Education Reform Act (KERA), the credentialing of teachers was moved from the SEA to the Education Professional Standards Board, an autonomous body appointed by the governor. KERA also established the Office of Education Accountability (OEA) under the Legislative Research Commission (LRC). The mission of OEA is to monitor the public education system and the implementation of KERA (CPRE, 1996). The State Board of Education (SBE) includes 11 members appointed by the governor and confirmed by the General Assembly and is responsible for managing the public schools, adopting policies for SEA, and hiring and evaluating the Commissioner of Education (CPRE, 1996).

## *Finance*

*How has proportion of state funding in education changed during this period?*
*How has expenditure per student changed during this period at the state level?*
*How has the gap between high- and low-poverty schools expenditures per student changed during this period?*
*To what extent were there efforts to equalize school funding between low-poverty and high-poverty districts during this period [questionnaire]?*
*To what extent were there efforts to link federal and state resources to support specific initiatives for poor/minority schools or students? Did the initiatives focused on mathematics and science learning [questionnaire]?*

;

*State Contribution to Education Funding*

Table A-1
*Trends in Proportions of State Contributions to Education Funding.*

| | 1987-88 | 1988-89 | 1989-90 | 1990-91 | 1991-92 | 1992-93 | 1993-94 | 1994-95 | 1995-96 | 1996-97 | 1997-98 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | %State | %State | %State | %State | %State | %State | %State | %State | %State | %State | %State |
| **Nation** | **49.54** | **47.79** | **47.11** | **47.16** | **46.37** | **45.6** | **45.2** | **46.8** | **47.5** | **48.0** | **48.4** |
| California | 70.20 | 66.44 | 66.86 | 66.03 | 65.87 | 62.2 | 56.2 | 54.2 | 55.8 | 60.0 | 60.2 |
| **Kentucky** | **65.19** | **68.06** | **67.34** | **66.95** | **67.02** | **67.0** | **65.9** | **65.8** | **65.3** | **62.9** | **61.7** |
| Maryland | 38.74 | 38.12 | 37.29 | 36.90 | 38.21 | 39.4 | 38.9 | 37.0 | 38.2 | 38.8 | 39.0 |
| Michigan | 35.28 | 32.53 | 26.81 | 26.77 | 26.57 | 30.6 | 28.7 | 67.3 | 66.8 | 65.5 | 66.0 |
| New York | 43.39 | 43.18 | 40.75 | 42.56 | 40.31 | 39.2 | 38.2 | 40.7 | 39.7 | 39.4 | 39.7 |
| North Carolina | 66.74 | 66.08 | 66.02 | 65.29 | 63.61 | 63.3 | 64.0 | 65.1 | 64.5 | 65.4 | 67.3 |
| South Carolina | 54.44 | 50.04 | 50.00 | 49.88 | 48.34 | 47.0 | 46.2 | 46.3 | 47.9 | 52.5 | 51.5 |
| Tennessee | 44.50 | 46.05 | 45.77 | 45.24 | 42.19 | 45.6 | 46.8 | 47.5 | 42.9 | 48.5 | 47.7 |
| Texas | 44.19 | 43.25 | 41.92 | 43.94 | 43.37 | 40.0 | 40.2 | 40.2 | 42.9 | 40.3 | 44.2 |
| Virginia | 32.62 | 33.84 | 32.69 | 32.75 | 31.11 | 32.1 | 30.8 | 31.8 | 31.1 | 32.5 | 31.4 |

Source: National Center for Education Statistics Research Department.

Table A-1 presents data on proportion of state contribution to the total education funding from school year (SY) 1988 to SY 1998. Across the years, Kentucky's average state contribution of 65.7% is higher than the national average of 47.2%. Over this ten year period, the proportion of state support ranged from a low of 61.7% in SY 1998 to a high of 68.1% in SY 1989. In comparison to the other states, Kentucky consistently had the highest state spending or close to highest state spending across the 10 years.

Table A-2
*Trends in Total Expenditures per Pupil.*

| | 1992-93 | 1993-94 | 1994-95 | 1995-96 | 1996-97 | 1997-98 |
|---|---|---|---|---|---|---|
| | Total ($) † | Total ($) † | Total ($) † | Total ($) † | Total ($) † | Total ($) † |
| **Nation** | **6,351** | **6,377** | **6,435** | **6,447** | **6,526** | **6,700** |
| California | 5,668 | 5,650 | 5,587 | 5,595 | 5,796 | 6,110 |
| **Kentucky** | **5,296** | **5,395** | **5,291** | **5,448** | **5,680** | **5,643** |
| Maryland | 7,445 | 7,414 | 7,482 | 7,472 | 7,443 | 7,615 |
| Michigan | 7,303 | 7,350 | 7,526 | 7,689 | 7,638 | 7,632 |
| New York | 9,545 | 9,662 | 9,675 | 9,475 | 9,393 | 9,583 |
| North Carolina | 5,437 | 5,437 | 5,475 | 5,348 | 5,431 | 5,691 |
| South Carolina | 5,165 | 5,191 | 5,240 | 5,416 | 5,564 | 5,759 |
| Tennessee | 4,514 | 4,566 | 4,676 | 4,728 | 5,048 | 5,344 |
| Texas | 5,246 | 5,374 | 5,563 | 5,685 | 5,803 | 5,893 |
| Virginia | 6,224 | 6,233 | 6,311 | 6,265 | 6,377 | 6,568 |

Source: National Center for Education Statistics. (1995, 1996, 1997, 1998, 1999, 2000). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1992-93, 1993-94, 1994-95, 1996-97, 1997-98*. Washington, DC: U.S. Printing Office.
† At constant prices of 2000-01

Table A-2 presents data on total expenditures per pupil for SY 1993 through SY 1998. Kentucky's average expenditure per student between SY 1993 to SY 1998 is $5,459, lower than the national average of $6,473. Compared to the other nine states, Kentucky spent less per pupil than most other states during this period.

Table A-3
*Variation in Instructional Expenditure per Pupil across School Districts*
*Gini Coefficient for instructional expenditures per pupil for unified districts, fiscal year 1988 to 1994.*

| | 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 |
|---|---|---|---|---|---|---|---|
| California | 0.075 | 0.083 | 0.088 | 0.087 | 0.085 | 0.074 | 0.047 |
| **Kentucky** | **0.093** | **0.098** | **0.084** | **0.070** | **0.075** | **0.075** | **0.078** |
| Maryland | 0.084 | 0.082 | 0.084 | 0.078 | 0.068 | 0.066 | 0.060 |
| Michigan | 0.098 | 0.099 | 0.104 | 0.105 | 0.094 | 0.091 | 0.085 |
| New York | 0.099 | 0.098 | 0.100 | 0.098 | 0.095 | 0.089 | 0.088 |
| North Carolina | 0.045 | 0.048 | 0.049 | 0.045 | 0.051 | 0.042 | 0.042 |
| South Carolina | 0.050 | 0.047 | 0.050 | 0.055 | 0.053 | 0.049 | 0.052 |
| Tennessee | 0.105 | 0.103 | 0.094 | 0.097 | 0.091 | 0.083 | 0.074 |
| Texas | 0.070 | 0.067 | 0.070 | 0.064 | 0.060 | 0.057 | 0.059 |
| Virginia | 0.104 | 0.105 | 0.091 | 0.089 | 0.089 | 0.079 | 0.081 |

Source: Hussar, W. & Sonnenberg, W. (2000). *Trends in disparities in school district level instructional expenditures per pupil* (NCES 2000-020) (p. 41).

Table A-3 presents data on heterogeneity in instructional expenditure per pupil across school districts. The tabled value of the Gini coefficient is a measure of how far the state was (in a given year) from uniform funding across school districts. A value of zero indicates perfect equality and larger values mean greater disparity.

Kentucky's average Gini coefficient is 0.082, higher than the 10-state average of 0.077, meaning that Kentucky has greater variability in instructional expenditures per pupil across their school districts than the 10-state average. Compared to the other nine states, Kentucky consistently fell near the middle in terms of the level of disparity. There is an indication that the disparities increased slightly after 1994.

## *Curriculum and Standards*

*To what extent was there a strong state curriculum in mathematics during this period [questionnaire]?*
*How strong were the mathematics content standards?*
*Was there a statewide textbook adoption in mathematics [questionnaire]?*
*To what extent were the mathematics standards linked to textbooks, performance standards and to assessment [questionnaire]?*
*How has this changed during this period [questionnaire]?*

### *Curriculum*

### *Standards*

Kentucky's standards in the four core subjects are described in the *Core Content for Assessment, which* was first developed in 1996 (AFT, 1996). By 1998, Kentucky developed *Program of Studies* and *Learning Descriptions* (at the elementary level) to clarify the standards (AFT, 1998). Starting in 1995, the American Federation of Teachers (AFT) reviewed all 50-states' standards and rated their curriculum standards. The math standards were rated as generally clear and specific about what students should know and be able to do at the elementary, middle, and high school levels (AFT, 1996; 1998). The science standards were rated as the strongest of the four subjects, offering clear and specific standards about the content students should learn at the elementary, middle, and high school levels (AFT, 1996, 1998). The English standards was rated as clear, specific and grounded in content at the elementary level; vague reading and basic writing conventions at the middle; and vague writing conventions at the high school levels (AFT, 1998). The social studies standards were rated as vague at the elementary and high school levels but clear at the middle school level (AFT, 1998).

## *Teacher Quality*

*Were middle grade certificates subject specific? When did this come into effect?*
*Was professional development tied to re-certification? When did this come into effect?*
*How have teacher compensation policies changed over time?*
*What was the extent of out-of-field teaching? How has this changed over time?*

Kentucky has a three-tiered certification system. Initial certification begins at Rank III, which requires a bachelor's degree and the completion of an internship. Rank II teachers must hold a master's degree or have completed a planned, fifth-year program. For fifth-year programs, professional development tied to an individual's professional growth plan may be substituted for up to 12 college credits, if approved by the college. Rank I requires 30 hours beyond the master's degree, acquired within eight years of attaining a BA. Recently, the state passed legislation allowing teachers to use National Board Certification as a substitute for graduate work to attain Rank I status (CPRE, 1996).

### *Initial Certificate*

Effective January 1985, teaching candidates must pass the state teacher certification test and complete a one-year internship before receiving a certificate that is good for 4 years. Thereafter, renewals are based on acquisition of a Master's degree, other advanced training and experience. Renewals are good for 5 year periods (Coley & Goertz, 1998; 1990).

Throughout the period 1988 to 1998, KY offered 3 types of middle grades (5-9) teaching certificates: 1) Provisional Certificate for Teaching in the Middle Grades 5-8, valid for 5 years and requiring a bachelor's degree; an approved program of preparation including student teaching and two middle grade teaching fields; and passing NTE or PRAXIS II scores (NASDTEC, 1988, 1991, 1996). This became the only certificate available in the middle grades 5-9 by 2000 (NASDTEC, 2000); 2) Standard Certificate for Teaching in the Middle Grades 5-8, which was discontinued on September 1, 1997. This certificate is valid for 5 years, requires eligibility for Provisional Certificate, and a master's degree or non-degree fifth-year approved program of preparation which includes a 12 semester-hour professional education component, and a 12 semester-hour specialization component (NASDTEC, 1988, 1991, 1996); and 3) Endorsement for Teaching in the Middle Grades 5-8, which was discontinued by 2000. This type of certificate allows elementary or high school certificates to be endorsed for teaching in the middle grades 5-8 upon completion of one middle grade teaching field and an approved program of professional preparation (NASDTEC, 1988, 1991, 1996, 2000). According to NASDTEC (1988, 1991, 1996) there is no requirement beyond the elementary credential for holders of elementary credentials to teach a departmentalized class in a specific subject in grades 7-9. While the requirements for holders of a secondary teaching credential to teach in a departmentalized class in a specific subject in grades 7-9 is a credential in the subject (where each teacher candidate must present at least one acceptable teaching major of 30 semester hours of credit or one area of concentration with 48 semester hours of credit. Teaching minors are accepted only in addition to an area or major???) (NASDTEC, 1988, 1991, 1996).

### *Professional Development*

Kentucky defined professional development as any course work, experience, training or renewal activity required by a state to keep a certificate in force (NASDTEC, 1996, 2000). Kentucky does not issue a permanent or life certificate. There are PD requirements to renew the second-stage certificate. The purpose of PD requirement is for continued employment (NASDTEC, 1996, 2000).

Kentucky requires that four days be set aside for teacher professional development and that each teacher receive at least 24 hours of training annually (Goertz, 1988; Coley & Goertz, 1990; CPRE, 1996). However, there is flexibility in how this requirement is met. Teachers may participate in school or district professional development during the school year, or they may satisfy the requirement on their own time, through summer coursework (CPRE, 1996). Legislation provides funding for one of the 4 days to be designated for a centralized or regionalized in-service program. In addition, funding is provided for the annual Commonwealth Institute of Teachers, a week-long seminar with follow-up weekends for up to 150-200 outstanding teachers (Goertz, 1988; Coley & Goertz, 1990).

*Teacher Compensation*

Table A-4 presents data on average teacher salaries from SY 1990 through 1997. Kentucky's average teacher salary, $37,729, was lower than the national average of $42,909. Compared to the other states, Kentucky's average teacher salary was the lowest in SY 1990 ($36,231) but fell in the middle ranges in the later years. However, Kentucky's average teacher salary is consistently lower than the national average across the years.

Table A-4
*Trends in Average Teacher Salaries.*

| | 1989-90† | 1990-91† | 1991-92† | 1992-93† | 1993-94† | 1994-95† | 1995-96† | 1996-97† |
|---|---|---|---|---|---|---|---|---|
| **Nation** | **43,180** | **42,991** | **43,344** | **43,125** | **42,886** | **42,774** | **42,661** | **42,309** |
| California | 51,881 | 51,147 | 51,214 | 49,044 | 48,661 | 47,340 | 47,782 | 47,369 |
| **Kentucky** | **36,231** | **38,068** | **39,122** | **38,225** | **37,888** | **37,568** | **37,488** | **37,243** |
| Maryland | 50,304 | 50,093 | 49,502 | 47,608 | 47,271 | 47,333 | 46,725 | 45,457 |
| Michigan | 50,229 | 49,424 | 51,563 | 51,912 | 54,149 | 54,217 | 53,752 | 52,631 |
| New York | 53,674 | 55,020 | 54,901 | 55,281 | 54,812 | 55,425 | 54,529 | 52,886 |
| North Carolina | 38,353 | 38,133 | 37,163 | 35,759 | 35,598 | 35,846 | 34,465 | 34,177 |
| South Carolina | 36,731 | 36,838 | 35,738 | 35,812 | 35,223 | 35,349 | 35,582 | 35,984 |
| Tennessee | 37,302 | 36,934 | 36,260 | 36,011 | 36,541 | 36,401 | 37,542 | 37,755 |
| Texas | 37,782 | 36,741 | 37,651 | 38,052 | 36,547 | 36,347 | 35,850 | 35,727 |
| Virginia | 42,644 | 42,745 | 40,849 | 40,413 | 40,083 | 39,471 | 39,311 | 39,793 |

Source: American Federation of Teacher Research Department, retrieved September 2003 from http://www.aft.org/research/salary/home.htm
† At constant 2000-01 dollars

*Out-of-Field Teaching*

Table A-5 presents data on the percent of 7-12th *classes* in the four core academic fields (math, English, science, and social studies) taught by teachers who did not have a minor or major in the field taught. The NCLB Act requires states to report in this manner for schools in terms of out-of-field teaching. Across the nation, the proportion of out-of-field classes increased from 21.8% to 24.2%. This pattern is also seen in Kentucky where the proportion of out-of-field classes increased from 29.5% to 31.7%. However, compared to the other nine states, Kentucky had the highest or close to the highest proportion of out-of-field classes in both years.

Table A-5

*Percentages of Public 7-12 Grade Classes in the 4 Core Academic Fields Taught by Teachers without a Major or Minor in the Field, by Year and State.*

| | 1993-94 | 1999-2000 |
|---|---|---|
| **United States** | **21.78** | **24.21** |
| California | 29.43 | 26.68 |
| **Kentucky** | **29.47** | **31.68** |
| Maryland | 20.71 | 22.33 |
| Michigan | 19.66 | 20.44 |
| New York | 13.12 | 18.11 |
| North Carolina | 17.72 | 19.41 |
| South Carolina | 23.54 | 22.49 |
| Tennessee | 29.52 | 35.62 |
| Texas | 20.95 | 29.67 |
| Virginia | 26.60 | 28.37 |

Source: Ingersoll, R. M. (2003). *Out-of-field teaching and the limits of teacher policy:* A research report (p. 19). Co-sponsored by Center for the Study of Teaching and Policy and The Consortium for Policy Research in Education.

## Assessment and Accountability

*How strong was the accountability system? What was the system of sanctions and rewards? How has this system changed over time [questionnaire]?*
*When was the mathematics assessment system first implemented? How has it changed during this period?*
*Was there a state-level mathematics assessment at the 8th or higher grades? What type was used (basic skills or higher order thinking)? Was it a requirement for graduation?*
*When did school report cards become available? When did the state start to disaggregate achievement data by subgroups [questionnaire]*

### *Accountability System*

The state was one of the earliest to gain approval for their accountability and assessment practices with the federal government (CPRE, 2000). The new emphasis of the accountability system has moved from rewards for teachers to rewards for school and from sanctions to assistance (CPRE, 2000). Carnoy & Loeb (2002) developed a zero-to-five index of the strength of accountability in 50 states based on the use of high-stakes testing to sanction and reward schools. Kentucky is of the a few states that received the second highest rating.

### *Student Accountability*

In 1996, Kentucky had no rewards or consequences for students linked to their standards (AFT, 1996). By 1998, although there were still no consequences for students who do not meet the standards, there were incentives for students to meet the standards. Students who meet or exceed credit requirements, which include Advanced Placement courses, and who maintain a “C” in all their classes earn the state's advanced "Commonwealth Diploma." (AFT, 1998). By 2000, student performance on the state assessment was sent to parents (CPRE, 2000).

*School Accountability*

*1998*. Kentucky provides funding for extra academic assistance to students who are having difficulty meeting the standards. Students are selected for the intervention based on teacher recommendation (AFT, 1998).

*2000*. The school accountability system is amended and will be implemented in 2000. Until then, the state has developed an interim model for charting school progress (CPRE, 2000).

*District Accountability*

*1996*. Kentucky law requires districts to provide "extended school services" to students who are not performing well enough to meet the state standards, and special funds are provided by the state for this purpose (AFT, 1996).

*2000*. A formal accountability system for districts has not been developed. However, a district summary of all of the district schools' report cards is printed in each area's largest circulation newspaper.

*State Assessment System (mathematics)*

*1988*. Students have been tested in reading, writing, Language arts, and mathematics in grades 3, 5, 7, and 10 since 1979. Local school districts must provide remedial help to those students falling below state-established performance standards. Under Legislation passed in 1984 the State Department of Education established essential skills standards by grade and subject in mathematics, reading, reference skills, spelling and writing and began testing students in these areas in every grade in 1985-86. Remediation is required for 1st and 2nd graders not passing the essential skills tests (Goertz, 1988).

*1990*. Students are tested in reading, writing, language arts, and mathematics in grades K, 1, 2, 3, 5, 7, and 10. Local school districts must provide remedial help to those students falling below state-established performance standards (Coley & Goertz, 1990).

*1996*. Kentucky has a state assessment system tied to their standards and given to all students across the state. Students are assessed in the core subject areas in grades 4/5, 7/8, and 11/12. The exact grade varies by subject (AFT, 1996).

*1998*. Beginning in the 1998/99 school year, Kentucky will implement a new testing program. According to state officials it will include reading and science tests in grades 4, 7, and high school; writing in grades 4, 7, and 12; and math and social studies tests in grades 5, 8, and high school (AFT, 1998).

*2000*. The state assessment system centers around the Commonwealth Accountability Testing System (CATS), the new testing system. CTBS-5 (is a multiple choice norm-referenced test) tests grades 3, 6, 9 in Reading, Math, Language Arts and CATS (is a multiple choice and open response criterion-referenced test) tests mathematics in grades 5 and 11 (CPRE, 2000).

## *Summary*

*Did the state have a consistent (in terms of duration and stability) commitment to education reform over the periods 1988 through 1998 [questionnaire]?*
*Did the state have a coherent approach (in terms of sequencing and alignment) to education reform during this period [questionnaire]?*
*What were the main reform mechanisms (curriculum control, teacher empowerment, standard-based assessment) during this period [questionnaire]?*
*What were the main positive effects of the reforms on low-poverty and high-poverty schools and students? What were the unintended effects [questionnaire]?*

## **References**

American Federation of Teachers. (1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997). *AFT 50-state teacher salary survey*. Retrieved as an excel spreadsheet on November 2003 from http://www.aft.org/research/salary/stgrave/Index.htm

American Federation of Teachers. (1995). *Making standards matter 1995: A fifty-state progress report on efforts to raise academic standards*. Washington, DC: Author.

American Federation of Teachers. (1996). *Making standards matter 1996: An annual fifty-state report on efforts to raise academic standards*. Washington, DC: Author.

American Federation of Teachers. (1998). *Making standards matter 1998: An annual fifty-state report on efforts to raise academic standards*. Washington, DC: Author.

Carnoy, M. & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Coley, R. J. and Goertz, M. E. (1990). *Educational standards in the 50 states: 1990*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Consortium for Policy Research in Education. (1996). *Teacher professional development profile: Kentucky*. Philadelphia, PA: University of Pennsylvania, Graduate School of Education, Consortium for Policy Research in Education.

Consortium for Policy Research in Education. (2000). *Assessment and Accountability in the Fifty States: Survey 2000: Kentucky Assessment and Accountability Profile*. Philadelphia, PA: University of Pennsylvania, Graduate School of Education, Consortium for Policy Research in Education.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved June, 2003 from http://epaa.asu.edu/epaa/v8n1/.

Goertz, M. E. (1988). *State educational standards in the 50 states: An update*. Princeton, NJ: Educational Testing Service.

Hussar, W. & Sonnenberg, W. (2000). *Trends in disparities in school district level expenditures per pupil* (NCES 2000-020). Washington, DC: National Center for Education Statistics.

Ingersoll, R. M. (2003). *Out-of-field teaching and the limits of teacher policy: A research report co-sponsored by Center for the Study of Teaching and Policy and the Consortium for Policy Research in Education*. Seattle, Washington: Center for the Study of Teaching and Policy.

Kentucky Department of Education. (2000). *State of Kentucky profile 1999-2000*. Lexington, KY: Author. Retrieved December 12, 2003, from http://www.kde.state.ky.us.

National Association of State Directors of Teacher Education and Certification. (1988). *Manual on certification and preparation of educational personnel in the United States.* Dubuque, Iowa: Kendall/Hunt Publishing Company.

National Association of State Directors of Teacher Education and Certification. (1991). *Manual on certification and preparation of educational personnel in the United States.* Dubuque, Iowa: Kendall/Hunt Publishing Company.

National Association of State Directors of Teacher Education and Certification. (1996). *Manual on certification and preparation of educational personnel in the United States.* Dubuque, Iowa: Kendall/Hunt Publishing Company.

National Association of State Directors of Teacher Education and Certification. (2000). *Manual on certification and preparation of educational personnel in the United States.* Dubuque, Iowa: Kendall/Hunt Publishing Company.

National Center for Education Statistics. (1993). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1992-93*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (1994). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1993-94*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (1995). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1994-95*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (1996). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1995-96*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (1997). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1996-97*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (1998). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1997-98*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (1999). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1998-99*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (2000). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 1999-2000*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

National Center for Education Statistics. (2001). *Statistics in brief: Revenues and expenditures for public elementary and secondary education: School year 2000-2001*. Retrieved December 7, 2003, from http://nces.ed.gov/ccd/pub_rev_exp.asp.

Rhoten, D., Carnoy, M., Chabran, M., & Elmore, R. (2003). The conditions and characteristics of assessment and accountability: The case of four states. In Martin Carnoy, Richard Elmore, and Leslie Santee Siskin (Eds.), *The new accountability: High schools and high-stakes testing* (pp. 13-53). New York, NY: Routledge Falmer.

## Appendix B: Questionnaire on State Education Policy

ETS Educational Testing Service

January 6, 2004

**Dear colleague,**

ETS is conducting a study on the relationship between state education policies and student performance on **NAEP 8th grade mathematics** from the late 1980s through the late 1990s, with a particular focus on changes in the Black-White achievement gap. States in the study are California, Kentucky, Maryland, Michigan, North Carolina, New York, South Carolina, Tennessee, Texas, and Virginia. You have been identified by our colleagues, as someone with considerable knowledge of **state's** policies during this period. We are asking you to assist us in three ways (listed below) so that we may complete our description of these policies. We expect this will take in all about two to three hours of your time.

Based on reports from a variety of sources such as the Council of Chief State School Officers (CCSSO), National Center for Education Statistics (NCES), *Education Week Quality Counts*, Education Commission of States, and Consortium for Policy Research in Education, we have compiled a profile on **state's** education policies over the period 1988 through 1998. We are concerned with the four major mechanisms by which states regulate, support and monitor the effectiveness of their education system: education finance and governance, curriculum and instructional policies, teacher quality, and accountability.

**First**, please read the profile and let us know if there are any errors or if we have omitted important information. **Second**, please fill out the attached questionnaire, which consists of questions that we were unable to address from secondary sources. Your responses to these questions are critical to our study as we hope to strengthen our understanding of the quality, coherence and consistency of state policy actions. **Third,** we will call within the next two weeks to schedule a 45-minute conversation with you to review the information from the questionnaire.

The data collected will be part of a policy report describing the relationship between states' policy actions and their success in raising the achievement of African American students and in closing the achievement gap with White students. We believe this study is unique in its attempt to develop a comprehensive, longitudinal description of both state policy histories and patterns of student achievement disaggregated by race and school-poverty level.

With the controversy surrounding the No Child Left Behind Act, it is all the more important that we examine our recent experience to glean some insights on what aspects of states' policies may account for their differential success in reducing achievement gaps. The report should be useful to policy makers as well as others interested in strategies for education reform and will be widely disseminated. With your permission, your contribution will be duly noted.

The profile and the questionnaire are provided in MS Word files. If you cannot retrieve the files or prefer to receive the questionnaire in another format, please contact us as soon as possible.

When you have completed the questionnaire, please send it as a MS Word attachment to hbraun@ets.org or awang@ets.org or fax it to (609) 734-5960.

If you have any questions about the study, the questionnaire, or the profile, please email or call Aubrey at (609-734-5058, awang@ets.org) or Henry at (609-734-5887, hbraun@ets.org). We will be happy to talk with you. We need your response to questionnaire by **January 27th**. We will also call to set up a 45-minute conversation with you to review the information from the questionnaire.

This study is being supported by the U.S. Department of Education. Unfortunately, we do not have funds to compensate you for your time. We do thank you in advance for your participation in this important policy study and will provide you with copies of the final report.

Sincerely yours,

| | |
|---|---|
| Aubrey Wang | Henry Braun |
| Associate Research Scientist | Distinguished Presidential Appointee |

## Questionnaire on State Education Policy

***Directions:*** *All questions should be considered in the context of state actions from the late 1980s through the late 1990s. We know that state policies evolve over time, and we are interested in understanding the development of those policies from about 1988 to 1998. If you have questions, feel free to call or email Aubrey Wang at ETS (awang@ets.org, 609-734-5058) or Henry Braun at (hbraun@ets.org, 609-734-5887). Thank you in advance for your help.*

I. The first series of questions focuses on the development of and changes in education finance policies in your state from 1988 through 1998.

I-1. How would you characterize the degree of equalization in school funding (between low-poverty and high-poverty districts) in the late 1980s?

I-2. How would you characterize the state's efforts to equalize funding through the 1990s?

I-3. Were there efforts during this period to link or combine state and federal funding (e.g., Title I; NSF State, Urban or Rural Systemic Initiatives; or Headstart) to support specific initiatives for poor/minority schools or students? If so, please describe.

I-4. Did these initiatives focus on mathematics and science learning? If so, please describe.

II. This series of questions focuses on the development of and changes in the mathematics curriculum and standards in your state from 1988 through 1998.

II-1. To what extent was there a state curriculum in 7th and 8th grade mathematics during this period?

II-2. How would you characterize the curriculum in terms of breadth, depth and rigor?

II-3. Was there a statewide textbook adoption policy in mathematics? If so, please describe (e.g., when did this occur, what was the policy).

II-4. Were the mathematics textbooks aligned with state curriculum? If so, please describe.

II-5. Were there mathematics content standards? If so, were the mathematics content standards linked to a matching set of performance standards during this period? If so, please describe.

III. These questions concern the changes and development in the accountability system in your state from 1988 through 1998.

III-1. When did school report cards become available?

III-2. When did the state start to disaggregate achievement data by subgroups?

III-3. Did non-Title I schools have school accountability during the period 1988 through 1998? When were these systems in place?

IV. Teacher Quality. (Please refer to the *Teacher Quality* section in the state profile for a summary of our understanding of this area of education reform).

**V. These questions concern the development of educational reforms in your state from 1988 through 1998**.

V-1. Who were the main drivers of education reform? What governmental and non-governmental entities drove education reform in this period? What kind of mechanism were in place to continue policy direction in the face of changing political environments?

V-2. What were the main reform mechanisms in your state (e.g., curriculum control, teacher empowerment, standards-based assessment) during this period? What is the evidence?

V-3. Based on your understanding of what occurred in the state during the period 1988 through 1998, do you think the state had a consistent (in terms of duration and stability) commitment to education reform through each of these mechanisms?

| Policy Area | Was there consistent commitment during this period? | Please describe what made this possible? | If these efforts were derailed, please describe the relevant factors or forces? |
|---|---|---|---|
| Education Finance | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |
| Curriculum and Standards | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |
| Accountability | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |
| Teacher Quality | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |

V-4. Based on your understanding of what occurred in the state during the period 1988 through 1998, do you think the state had a <u>coherent approach</u> (in terms of sequencing and alignment of the reform) to education reform?

| Policy Area | Was there coherence to the education reforms during this period? | Please describe what made this possible? | If these efforts were derailed, please describe the relevant factors or forces? |
|---|---|---|---|
| Education Finance | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |
| Curriculum and Standards | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |
| Accountability | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |
| Teacher Quality | Yes [ ]<br>Somewhat [ ]<br>No [ ] | | |

V-5. Overall, what were the (generally agreed upon) positive effects of the reforms? What were the unintended effects?

VI. Finally, have we missed anything? Are there other matters that you would like to bring to our attention? Are there references you could recommend? Who else should we speak to? Please provide any comments or information on any other topics related to state policy actions in your state during this period that you feel are important.

Thank you very much for your assistance.

# Appendix C: Follow-Up Interview Protocol

## Follow-up Questions for State Experts (Example for Kentucky)

I-1. When did we see influx of state funds into education (our table doesn't show it)?

I-2. Were these state initiatives focused on disadvantaged minorities in particular, for instance, programs focusing at-schools or mathematics and science?

I-3. NSF SSI-RSI, any information?

II-1. Was the link between content standards and performance standards clear enough to inform teachers and were there TPD support for improving pedagogy? Standards for all? Teaching?

III-1. School report cards?

IV. Any special initiatives on TPD that are related to reform?

V. Books on KY: Accountability consistency, assessment-major changes, coherence and timing—especially accountability leading other reform components

VI. KY specific issues on demographics of poverty

# Appendix D: Example of a State Summary

## *Kentucky State Summary*

I. Governance and Politics in the Context of the Reform

A. State governance: Degree of central vs. local control

**Essential Points**

- State in charge of assessment, accountability
- Localities in charge of curriculum, professional development

**Summary**

Governance of education reform in Kentucky was divided between the state department of education and the schools. The state department of education created and managed the assessment and accountability systems, while the schools and school-based decision making councils decided on the curriculum and resource distribution.

B. Main drivers of education reform

**Essential Points**

- State Supreme Court case and decision – brought by grassroots non-governmental organizations together with business organizations and a coalition of districts
- Supported by state Board, Commissioner of Education, Governor, and General Assembly
- Succession of leaders of the Kentucky Teachers Association supported the reform when they were promised that the reform would not touch the evaluation of teachers.
- During the decade of the 90's Kentucky changed governors, changed Commissioners of Education, and changed crucial Department of Education personnel. Despite the changes, there was a sense of continuity about the reform. New governors and commissioners embraced the reform as ardently as did their predecessors. Several State Board of Education members held their positions and continued to support the reform during the being studied.

**Summary**

The main drivers of Kentucky's educational reform during the late 1980s through the 1990s were a combination of local coalition of grassroots non-governmental organizations such as the Prichard Committee, the Forward in the Fifth, together with business organizations and a coalition of districts who brought the lawsuit that resulted in the landmark state Supreme Court ruling that the state's public school system was unconstitutional. As a result of this court ruling, the General Assembly drafted the blue print of Kentucky education reform: the Kentucky Education Reform Act (KERA) in 1990, which was amended and sustained through a combination of the election of state board and the commissioner and continuous support by the General Assembly. Though the central thrust of KERA was targeted at fiscal equity, it set the stage for a host of wide-ranging reforms.

First state reform to follow the Governor's Conference in 1989 and enact those recommendations.

C. Main reform mechanisms

**Essential Points**

- Finance reform
- Standards-based assessments
- School based accountability
- Increasingly explicitly/rigorously defined standards

**Summary**

The main reform mechanisms in Kentucky were extensive finance reform, the promulgation of statewide standards, standards-based assessment, and school-based accountability. Passage of the Kentucky Educational Reform Act (KERA) required a number of dramatic changes in state oversight of public education. Primary among these was the restructuring of school finances to create a more equitable system. This included the addition of a significant amount of money.

KERA also increased the state role dramatically in standards setting, assessment and accountability. Though Kentucky will not adopt statewide curricula due to state law, it has created a set of standards that between 1990 and 1998 became increasingly well defined in response to public demand. Because these standards were assessed through a number of performance measures, teachers and administrators demanded clearer alignment, thereby increasing the state role. Additionally, rewards and sanctions, delivered by the state, were attached to performance for the first time with the passage of KERA.

Furthermore, the legislature established an Office of Educational Accountability (OEA) that reported to the Legislature about educational reform. The OEA was concerned about the quality of the assessment and accountability system and did extensive work, such as investigating allegations of cheating and conducting/sponsoring evaluations of the state department of education and its contractors.

II. Finance

A. Continuous commitment in education finance

Essential Points

- Prodded into action by a court decision in 1989, the Legislature has maintained a commitment to providing equitable and adequate school funding.
- There has not been an overwhelming growth in per pupil funding and the state remains well below the national average.

Summary

The financial inequity among schools within the state, more than the overall funding level, was the impetus for education finance reform in Kentucky. In 1989, the state supreme court decided that Kentucky's system of public education did not fulfill its constitutional obligation. As a result, the state went through a major overhaul of the system, of which finance was a large part. Traditionally, districts in the urban centers have spent considerably more money than those in the rural areas and the eastern part of the state. According to all reports, efforts at funding equalization have received consistent attention from the state legislature as well as from local communities over the time period being considered here.

While Kentucky has tried to equalize funding through a shared state and local commitment, the total expenditures per pupil have not risen dramatically. This is due in part to significant expenditures at the state level in redesigning the system (additional state spending is not represented in per pupil averages, leading to the question about finance discrepancies that I raised earlier) and in part to a desire to redistribute spending rather than raise the total amount.

B. Level and trajectory of proportion of state funding in education

Essential Points

- Per pupil spending remains well below the national average.
- No clear trajectory in per pupil spending in the state from 1992 to 1998 (possible that the big jump took place in 1990 with passage of KERA).

Summary

The percentage of total educational expenditures paid for by the state reached a high of 68% in 1988 and has been dropping since then (with one small exception). The size of the drops began to increase in 1995, ending with just 61.7% of expenditures paid for by the state in 1997. The average expenditure per student does not follow the fluctuation in state contribution, indicating fluctuation in the levels of local and federal contributions. Overall, Kentucky's per student spending is considerably below the national average for the period from 1992-1998. Its expenditures average $5,459 and range between $5,291 and $5,680, without any clear trajectory in either direction. In contrast, the national average of per pupil expenditures for this period was $6,473 and demonstrated a clear and continuous annual increase during the six years.

C. Level and trajectory of gap between high- and low-poverty school expenditures per student

Essential Points

- Passage of KERA reduced inequities in funding in 1990.
- Inequities have been increasing since 1992 and remain above the average for our ten states.

Summary

Kentucky's average Gini coefficient is 0.082, higher than the 10-state average of 0.077, meaning that Kentucky has greater inequity in instructional expenditures per pupil across their school districts than the 10-state average. Compared to the other nine states, Kentucky consistently fell near the middle in terms of the level of disparity. After improving on this measure of equity in 1990 and again in 1991, inequity appears to have been creeping back into the system since 1992.

D. Characterization of equalizing efforts in school funding

Essential Points

- After initial improvement in equalizing school funding, the state has lost ground.
- Many respondents reported that the state has been committed and successful in this area, though our data does not necessarily support this contention.

Summary

As stated, Kentucky's efforts at equalizing school funding were required by a 1989 court decision. In 1990, the state legislature passed the Support Excellence in Education in Kentucky funding program. This program sought to use a formula-driven funding program to meet district needs on an equitable basis. In order to fund the equalization, the state increased the sales tax from 5 cents to 6 cents. At least initially, these efforts resulted in greater equity among schools. However, as stated above, the inequity among schools has been on the rise after that initial correction.

III. Curriculum and Standards

A. Continuous commitment in curriculum and standards

Essential Points

- State was initially very vague about curriculum and standards and only achieved widespread core content standards at the end of the time period we are studying.
- Reluctance to mandate any state curriculum or materials has lead to widely varying implementation of curriculum.

Summary

Curriculum only became a state focus with the passage of the 1990 legislation. However, state law prohibits the establishment of a statewide curriculum. As a result, the state embarked upon establishing a set of state standards in math. Since 1990, the state has demonstrated a strong commitment to a standards-based approach and has shown a commitment to developing increasingly detailed state content standards upon which assessment is based. This resulted in core content standards in 1998. The staff at the Kentucky Department of Education attempts to make standards clear without prescribing curriculum.

The establishment of standards is somewhat further complicated by the fact that textbook selection is decided at the school level. In combination with evolving standards, it is unclear to what extent the standards and assessments moved school staff toward a common curriculum in the state. For example, discussions continued during much of the time period in question about the extent to which algebra should be taught in grades 7, 8, or 9. As one respondent indicated, while standards documents now exist in some detail, schools in his district continue to pursue very different curricula based on textbook selection, tracking practices, and instructional decisions and resources. This evidence would indicate that while the state has demonstrated an evolving commitment to standards, the impact that this has had at the local level may be quite variable.

B. Strong state curriculum in mathematics (with statewide textbook adoption, alignment of textbook with curriculum, and content standards linked to performance standards)

Essential Points

- The state has a list of approved texts but schools make their own choices.
- Curriculum alignment was one of the state's criteria for selecting texts but they materials represent a wide range of approaches and even content.
- The state did create content standards but did not provide clear enough guidance for teachers, nor a clear sequence of progression.

Summary

As stated above, the state standards went though a process of evolution. Their strength is a subject of some debate. For example, the 56 "Valued Outcomes" that the state identified were elaborated through two or three versions, but the lists always consisted of "examples," not a definitive "scope," and there was never any sequencing laid out by the state. That was in keeping with the idea that Kentucky was a "local control" state with regard to curriculum.

Kentucky did provide a list of approved textbooks (from which districts could generate their own "approved list") and schools were able to choose and provided state funds for textbook adoption. However, the list of approved texts represented a broad range of instructional and curricular approaches. There were as many as five different choices in each content area by grade. While one of the criteria for state approval was alignment with standards, the range of approved texts raises some questions about degree of alignment. All respondents agree that the state and text publishers would argue that the texts were aligned with standards, but the standards were quite vague at the start and have become only progressively more clear while new approved lists were not always created concomitant with the creation of new standards.

The content standards that the state developed were general descriptions of content and performance. In math, such content standards were not developed until the mid-1990's and were not strongly linked to performance standards until 2001. They included examples of tasks/problems students should be able to do at each grade level.

Beginning in 1998, the state made clear the essential content for all students to know that would be included in the state assessment. This was to be used in conjunction with previously released Academic Expectations and Programs of Studies. Prior to this, there had been some sharing of assessment items rather than curriculum.

C. Breadth, depth and rigor of mathematics curriculum

Essential Points

- Math curriculum was not made explicit until the mid to late 1990's.
- AFT describes math standards as clear and specific, though this is not an evaluation of rigor.

Summary

Kentucky's standards in the four core subjects are described in the *Core Content for Assessment, which* was first developed in 1996 (AFT, 1996). By 1998, Kentucky developed *Program of Studies* and *Learning Descriptions* (at the elementary level) to clarify the standards (AFT, 1998). Starting in 1995, the American Federation of Teachers (AFT) reviewed all 50-states' standards and rated their curriculum standards. The math standards were rated as generally clear and specific about what students should know and be able to do at the elementary, middle, and high school levels (AFT, 1996; 1998). Kifer would disagree, feeling that the 8th grade math standards are nearly incoherent and do not help teachers to cover essential understandings.

IV. Teacher Quality

A. Continuous commitment in teacher quality initiatives

Essential Points

- State did increase funding to be used for professional development; use was a local decision.
- State began to provide content-based assistance to teachers in 1992.

Summary

The state did establish the Professional Standards Board in order to strengthen teacher quality. However, the Board did not have a great deal of authority and had little ability to deal with recruitment or retention of high quality teachers – this remained a local responsibility. Districts with strong teacher associations were able to supercede state initiatives that would have impacted the teacher requirements. Additionally, the state's school accountability system did not include teacher quality as a variable in determining school ratings and thus did not receive a great deal of direct attention.

However, starting in 1990, the state did provide money for teacher professional development. There was local flexibility on how the time and money should be used but it was meant to build teachers content base. Starting in approximately 1992, the state began to provide direct content-based assistance to teachers. Additionally, distinguished educators in the schools (so designated by the state) had the power to terminate teachers who were documented as chronically performing below expectations. The 1990 legislation also raised teacher salary, in the hopes that competition would increase in under-served areas.

B. Middle grades content specific teacher certification, especially in mathematics

Essential Points

- Kentucky used to have less rigorous standards for middle grade certification in particular subjects.
- By 2000, the state began to require newly and re-certified teachers to have passed a national teaching exam and have two field specialties.
- Within five years, all middle grade teachers should be required to have this higher level certification.

Summary

Effective January 1985, all teachers in Kentucky must pass the state teacher certification test and complete a one-year internship before receiving a certificate that is good for 4 years. Thereafter, renewals are based on acquisition of a Master's degree, other advanced training and experience. Renewals are good for 5 year periods (Coley & Goertz, 1998; 1990).

Throughout the period 1988 to 1998, KY offered 3 types of middle grades (5-9) teaching certificates. Only the first of these still remains as an option for middle grade teachers:

1) Provisional Certificate for Teaching in the Middle Grades 5-8, valid for 5 years and requiring a bachelor's degree; an approved program of preparation including student teaching and two middle grade teaching fields; and passing NTE or PRAXIS II scores (NASDTEC, 1988, 1991, 1996). This became the only certificate available in the middle grades 5-9 by 2000 (NASDTEC, 2000).

2) Standard Certificate for Teaching in the Middle Grades 5-8, which was discontinued on September 1, 1997. This certificate is valid for 5 years, requires eligibility for Provisional Certificate, and a master's degree or non-degree fifth-year approved program of preparation which includes a 12 semester-hour professional education component, and a 12 semester-hour specialization component (NASDTEC, 1988, 1991, 1996).

3) Endorsement for Teaching in the Middle Grades 5-8, which was discontinued by 2000. This type of certificate allows elementary or high school certificates to be endorsed for teaching in the middle grades 5-8 upon completion of one middle grade teaching field and an approved program of professional preparation (NASDTEC, 1988, 1991, 1996, 2000). According to NASDTEC (1988, 1991, 1996) there is no requirement beyond the elementary credential for holders of elementary credentials to teach a departmentalized class in a specific subject in grades 7-9. While the requirements for holders of a secondary teaching credential to teach in a departmentalized class in a specific subject in grades 7-9 is a credential in the subject (where each teacher candidate must present at least one acceptable teaching major of 30 semester hours of credit or one area of concentration with 48 semester hours of credit. Teaching minors are accepted only in addition to an area or major.) (NASDTEC, 1988, 1991, 1996).

C. Professional development tied to re-certification

Essential Points

- A minimum of 24 hours per year of professional development is required for all teachers.
- Increases in teacher rank require attainment of a master's degree (Rank 2) or a master's degree plus 30 credit hours (Rank 1).

Summary

Kentucky defined professional development as any course work, experience, training or renewal activity required by a state to keep a certificate in force (NASDTEC, 1996, 2000). Kentucky does not issue a permanent or life certificate. There are professional requirements to renew teaching certificates and to advance in teaching class (NASDTEC, 1996, 2000).

Kentucky requires that four days be set aside for teacher professional development and that each teacher receive at least 24 hours of training annually (Goertz, 1988; Coley & Goertz, 1990; CPRE, 1996). However, there is flexibility in how this requirement is met. Teachers may participate in school or district professional development during the school year, or they may satisfy the requirement on their own time, through summer coursework (CPRE, 1996). Legislation provides funding for one of the 4 days to be designated for a centralized or regionalized in-service program. In addition, funding is provided for the annual Commonwealth Institute of Teachers, a week-long seminar with follow-up weekends for up to 150-200 outstanding teachers (Goertz, 1988; Coley & Goertz, 1990).

D. Level and trajectory of teacher compensation

Essential Points

- In 1989, Kentucky's average salary was the lowest of the ten states being studied.
- With the passage of KERA, average salaries increased for two years before beginning a steady decline.
- By 1997, Kentucky's average salary still remained below the national average but had moved closer to the middle of the ten states being studied here.

Summary

Kentucky's average teacher salary between 1989 and 1997, was lower than the national average during that same time period ($37,729 vs. $42,909). Compared to the other states, Kentucky's average teacher salary was the lowest in SY 1989 ($36,231). Clearly, the salary received a significant increase with the passage of new legislation in 1990, jumping almost $2,000 on average, and another $1,000 the following year. This trend toward higher salaries lasted only two years and since 1992 average salary has been declining in Kentucky. However, the state falls in the middle range of the ten states in the later years, as some states saw near continuous declines in average salary.

E. Extent of out-of-field teaching

Essential Points

- Like most other states as well as the nation as a whole, the proportion of classes being taught by teachers not majoring or minoring in the field grew in Kentucky between 1993 and 1999.
- Kentucky consistently has the second highest proportion of classes with out-of-field teachers in our set of ten states.

Summary

Across the nation, the proportion of middle and high school classes in the four core academic fields (math, English, science, and social studies) being taught by teachers who did not have a minor or major in the field increased from 21.8% to 24.2% between 1993 and 1999. This pattern is also seen in Kentucky where the proportion of classes being taught by out-of-field teachers increased from 29.5% to 31.7%. Compared to the other states we are studying, Kentucky had the second highest proportion of classes (ranking just above Tennessee) being taught by out-of-field teachers in both years.

V. Assessment and Accountability

A. Continuous commitment in assessment

Essential Point

- Assessment has been continuous though grades, subjects, and modes of assessment have varied over the years.

Summary

Though the state has demonstrated a clear commitment to assessing students since the passage of KERA in 1990, there has been evolution in the mode of assessment. From 1988 through 1991, the state used a norm-referenced test. From 1992-1998 the state used a performance-based and open-response portfolio assessment system. From 1998 until present, the state uses a combination of the performance-based assessment and a norm-referenced test.

B. Type of assessment, grade level assessed, requirement for graduation

Essential Point

- From 1990-1998, students were assessed in grades 4/5, 7/8, and 11/12 (exact grade depended on which of the four core content areas was being measured) through a rich and time-consuming performance based assessment system.

Summary

Since 1979, Students have been tested in reading, writing, language arts, and mathematics in grades 3, 5, 7, and 10 (K, 1, and 2 were added in 1990). Local school districts must provide remedial help to those students falling below state-established performance standards (Coley & Goertz, 1990). Between 1990 and 1998, Kentucky had a state assessment system tied to its standards and given to all students across the state. Students were assessed using performance measures including in-class activities and portfolio assessments the core subject areas in grades 4/5, 7/8, and 11/12. The exact grade varied by subject (AFT, 1996). Beginning in the 1998 school year, Kentucky implemented a new testing program. It included both multiple choice and open-ended response. The state uses the CTBS-5 (a multiple choice norm-referenced test) in grades 3, 6, 9 in reading, math, language arts. In grade 5 and 11, the state uses CATS (a multiple choice and open response criterion-referenced test) to assess mathematics. (I do not have information on what may have been required for graduation during the time period of our study.)

C. Consistent commitment in accountability

Essential Point

- Consistent commitment to holding "someone" responsible for student performance.
- The "someone" has shifted between the school/teachers and the district, not directly impacting the students.

Summary

Though the state has clearly shown a consistent commitment to a system for educational performance accountability, the focus of the state accountability system did shift during the time period under consideration. From 1988 to 1992, the system focused on accountability for whole school districts. From 1993 to 1998, accountability was much more centered at the school level. The system has continued to develop since 1993, adding both indicators and incentives. For example, rewards and sanctions were applied to teachers beginning in 1993 or 1994. Between 1988 and 1998, there was no state mandated individual student accountability in Kentucky.

D. Strong accountability system (with an effective system of sanctions and rewards) for both Title I and non-Title I schools?

Essential Points

- The strength of the accountability system has been increasing since KERA was passed.
- It moved from a district focus (which allows individual schools to "hide"), to a school and teacher focus during the time period we are studying.

Summary

All public schools in the state have been included in the accountability system since 1992, when accountability was established at the individual school (rather than district) level. The strength of the system has been increasing as elements (such as rewards and sanctions for teachers) are added. Currently, the system receives a 4 out of 5 on a scale of accountability strength developed by Carnoy and Loeb (2004).

VI. Overall Quality of State Reform or "State Story".

A. *Alignment in policy*

- Continuous improvement in alignment of standards and assessment.

Kentucky's system of public education underwent an enormous upheaval with the Kentucky Education Reform Act in 1990. This was the first state to engage in the comprehensive reform along the lines of what was recommended in 1989 at the Governors' Conference. The system that the legislation created attempted to improve finance, governance, assessment, accountability, and teacher training. However, without a clearly defined set of statewide learning objectives, something that did not exist in 1990, it was difficult to say just what this alignment was around. The state increasingly paid attention to curriculum standards as the difficulties of alignment became clear. Nonetheless, the state did create a statewide performance based assessment system and was able to demonstrate progress on a number of measures. Kentucky started with the assessments and had the standards follow, rather than vice versa. The order was not ideal, and they had to do more revision than they would have, had they had the patience to allow assessment and accountability lag behind the other reforms.

B. *Consistency in policy*

- <u>*Consistent in terms of goals, fluctuation in terms of methods.*</u>

While Kentucky has maintained a consistent statewide focus on improving educational outcomes, and has committed enormous resources to this task, the areas of state policy emphasis have shifted over the years. For example, the state began to provide increasingly explicit curricula and standards due in large part to the demands of teachers around the state. Similarly, the state scaled back the in-depth performance assessments that had been created initially because of the popular discontent with the large about of time that was being used for the purpose of formal statewide assessments.

Perhaps the most consistent policy change has been the modification in state education finance. Though we have some discrepant information on the amounts of state contribution and its increase, all agree that the ways in which schools were financed changed dramatically in 1990 and has been relatively stable since then.

C. *Quality of policy implementation*

- In the areas of standards and curriculum, local authority has resulted in varying curricula and content pacing.

While state policy goals related to finance and governance can be easily accomplished by legislative fiat, implementation of curriculum and standards is more difficult. This was made more difficult in Kentucky because of the way in which this material evolved. Core content was not made clear until 1996 and even this has to be made more explicit for teachers with a Program of Studies document that was published in 1998. In addition, the multiple iterations of the assessment system, from norm referenced traditional assessment in 1988 to performance based criteria referenced in 1993, served as a significant change in the implementation of the assessment system.

Despite these changes, it should be noted that the changes to the state's education system were most dramatic in 1990. Though there has certainly been modification since then, all respondents report a seismic shift in the way the state was running the public education system.

D. *Perceived positive effects*

- Change in attitude and focus
- More shared language around educational goals
- Increased supplementary and support programs

The fundamental shift in education policy that occurred following the court decision had a number of positive consequences for the state. Primary among them was a change in attitude among educators and the general public. Raising student achievement became a primary goal. The system was intentionally designed to give schools the most credit for raising the lowest performers to a higher level. This put a particular emphasis on improving educational outcomes for struggling students. This was the first time that the state began to talk about including all students in the assessment process. In so doing, it raised the status of previously under-served populations including those with disabilities or particularly low test scores. When the statewide data that demonstrated that many poor schools were making significant improvements in performance convinced some that poor students could learn, it undermined a previous feeling that student demographics and school context were insurmountable barriers.

With these reforms came nationwide positive attention. Actors in the state and local systems felt that the state had taken a very positive step toward improving education and thus were more convinced of its potential for success. Additionally, the reduction in variation of school funding helped many to be more comfortable with the direction of change. The changes reduced the place of education as a wedge issue dividing policymakers. There was widespread support for many of the reforms, reinforced by the national attention they garnered.

The reforms provided teachers and administrators across the state with a common language and set of goals (though they were initially vague) to discuss their own and their students' work. This occurred in terms of classroom instruction but also in terms of larger issues related to testing, assessment, and the particular needs of individual students. For example, Kentucky's district assessment coordinators, a position created by the 1990 legislation, created their own professional organization, which has been a powerful group for professional development and advocacy. The assessment system pushed many in the state to focus on higher order thinking skills and applications as opposed to the more basic-skills approach that had existed prior to the 1990 legislation.

Kentucky also created a system of parallel supports in its 1990 legislation. More social supports for students and families, early childhood education, increased assistance to struggling students, were all additional components of the reform that resulted in a better articulated system of services for both students and families in the state.

E. *Perceived unintended/negative effects*

- Significant early fluctuation in policy, particularly in the early years.
- Shifts in emphasis due to accountability pressures.

Because of its dramatic changes and its early steps in the accountability movement, Kentucky received a great deal of national and research attention. As mentioned above, when this attention was positive, it served a great purpose in the state. However, because the state was in such a period of policy transition, the early research had dramatic effects on the direction of policy in sometimes unintended ways. The effects of external evaluations of the system also served to undercut the credibility of the system in some ways.

Of course, Kentucky has not been insulated from the challenges posed by all accountability systems. For example, when the math portfolio was criticized by math teachers as being too heavily focused on writing, it was eliminated after being in place for three years. Other portfolios, intended to be used to improve instruction have become classes unto themselves (students call it "portfolio prison"), demonstrating that even more performance based forms of assessment are subject to the dangers of overly prescriptive test preparation.

The accountability system also made teachers nervous at first. When rewards and sanctions began to be targeted at the school level, teachers became quite nervous about their job security. At those schools that merited rewards under the system, the local decision of how to distribute the award in some cases undercut the types of cooperation and collaboration that the system was hoping to encourage.

F. *Lessons learned*

Kentucky continues to refine its systems of standards, assessment, and accountability. The state is still not strong in terms of student accountability. In spite of this, many believe that the emphasis on performance and accountability has perverted instruction in some ways, shifting attention to particular areas of assessment rather than a more balanced and consistent approach.

G. *Special characterization of the state*

Kentucky attempted to address a huge range of issues simultaneously. Initiatives focused on all core subjects and all grade levels and included finance, governance, curriculum, assessment, and accountability. They even attempted to institute a range of support programs for children and families at the same time. Being both earlier and more far-reaching than many states posed its own set of challenges and information gaps with which the state had to contend while it was undergoing a massive systemic adjustment.

Demographic information – poor minorities are in the urban areas, while poor white kids are in the rural areas so there is limited mixing of poor students due to large scale geographic shifts.

# Appendix E: HLM Variables

*Student Variables*

*Socioeconomic Status (SES)*. SES is the average of non-missing values of the following two variables.

1. Number of reading related items in the home (i.e., Newspaper, encyclopedia, books, magazines)
2. Educational level of parent with greatest education.

If both variables are missing, SES is defined as missing. If exactly one variable is non-missing, SES consists of the non-missing variable. The SES scale has been standardized within state to have mean 0 and unit standard deviation.

*Academic focus for the student (AcadFoc)*. AcadFoc is the sum of the non-missing values of a set of variables, displayed in Table E-1, that are hypothesized to measure the construct of student academic focus. Results of a confirmatory cluster analysis were consistent with the hypothesis. Note that six variables occur in all years, five occur in two adjacent years and one variable only occurs in 1992. (Note: All variables available for a given year were used in the analyses for that year.) The scale of the composite variable was standardized within state to have zero mean 0 and unit standard deviation.

*Black vs. White (BvsW).* This is an indicator variable that takes the value zero if the student is White, 1 if the student is Black and missing if the student is Hispanic, Asian, Pacific Islander. If a school has no Black or White students, this variable is missing for the entire school. In the level 1 model, the regression coefficient associated with this variable is equal to the mean difference between Black and White students for the school or the stratum, depending on the structure of the corresponding level 2 model.

*Average school SES (AggSeS).* The average of the SES values of students in the NAEP sample in the school.

*Percent Black Students Assessed in the School (AggBvsW)*. The average of the Black/White indicator in the NAEP sample in the school. This is equal to the proportion of Black students in the subset of Black or White students in the sample.

*School Climate (Climate)*. The average of the non-missing values of those questionnaire items having to do with school climate, based on responses by teachers and school administrators.

Table E-1 presents a brief description of each variable and indicates the administrations in which they were measured.

Table E-1
*Student Variables Included in the Academic Focus Composite.*

| Variable | Description | 92 | 96 | 00 |
|---|---|---|---|---|
| Homework | B003901 How much time is spent each day on homework.<br>(This was recoded to collapse the 1st 2 categories.) | X | X | X |
| Pgsread | B001101 How many pages read in school and for homework.<br>Recoded to reverse the categories. | X | X | X |
| Textbk | M811601- How often? Do math problems from textbooks.<br>Recoded to reverse the categories. | X | X | X |
| Calc | M811605 How often? Use a calculator.<br>Recoded to reverse the categories. | X | X | X |
| Mathhi | IF Classnow =3 or 4, Mathhi=1. Otherwise mathhi=0.<br>Classnow- M810501 What kind of class are you taking this year? | X | X | X |
| Mathhi9 | If class9 =4 or 5, mathhi9=1. Otherwise mathhi9=0.<br>Class9- M811701 What math class will you take in 9th grade? | X | X | X |
| Wrksheet | M811602 How often? Do math problems on worksheets? | X | X | |
| Reports | M811611 How often? Write reports or do math projects. | X | X | |
| Solution | MM00501 How often asked to provide detailed solutions? | X | X | |
| Tests | M811607 How often? Take math tests. (same categories as M811602). | X | X | |
| Howfar | B009801 How far in school do you think you will go?<br>Recoded so that category 6=1. | | X | X |
| Makeup | M811610 How often? Make up problems for others to solve. | X | | |

### *School Variables*

*Aggregated Academic Focus (AggAcadFoc).* The average of the student academic focus variable for the students in the NAEP sample in the school.

The following table indicates the variables comprising this construct and the administrations in which each was measured. Due to changes in the background variables, as well as differences in how well the variables related to the construct, a different set of variables was used for each administration. Of course, this limits comparability across administrations.

Table E-2

*School variables included in the school climate construct.*

| Variable | Description | 92 | 96 | 00 |
|---|---|---|---|---|
| Absenteeism Student | Is student absenteeism a problem in your school? | X | X | X |
| Absenteeism Teacher | Is teacher absenteeism a problem in your school? | X | X | X |
| Fighting | Are physical conflicts a problem in your school? | X | X | X |
| Parent Support | Is parent support for academics positive or negative? | X | X | X |
| Pct Absent | What percent of students are absent on an average day? | X | X | X |
| Pct Retained | What percent of students are retained in this grade? | X | X | X |
| Pct Enroll | What percent of students are enrolled at the beginning and the end of the year? | X | X | X |
| Tardiness | Is student tardiness a problem in your school? | X | X | |
| Pct PTA | What percent of parents are in a parent-teacher organization? | | X | X |
| Pct Par Schl | What percent of parents attend open house/back-to-school night? | | X | X |
| Par Involve | Is lack of parent involvement a problem in your school? | | X | X |
| Gangs | Are gang activities a problem in your school? | | X | X |
| Misbehavior | Is student misbehavior a problem in your school? | | X | X |
| Pct Leave | What percent of full-time teachers leave before the end of the year? | | X | X |
| Cut Class | Is cutting classes a problem in your school? | X | | |
| Teacher Acad | Are teacher attitudes to academics positive or negative? | X | | |
| Property | Is regard for school property positive or negative? | X | | |
| Teacher & Student | Are relations between teachers and students positive or negative? | X | | |
| Pct Conf | What percent of parents attend parent/teacher conferences? | | X | |
| Health | Is student health a problem in your school? | | X | |
| Cheating | Is student cheating a problem in your school? | | X | |
| Expectations | Are teacher's expectations for student achievement positive or negative? | | | X |

Seven variables are common to all three years, another seven are common to two adjacent years and eight variables occur in only one year.

**EDUCATION POLICY ANALYSIS ARCHIVES** **http://epaa.asu.edu**

**Editor: Sherman Dorn, University of South Florida**

***Production Assistant: Chris Murrell, Arizona State University***

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Sherman Dorn, epaa-editor@shermandorn.com.