DOCUMENT RESUME

ED 404 366 TM 026 451

AUTHOR Berends, Mark; Koretz, Daniel
TITLE Reporting Minority Students' Test Scores: How Well Can the NAEP Account for Differences in Social Context?
INSTITUTION National Assessment of Educational Progress, Princeton, NJ.; Rand Corp., Washington, D.C.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE 96
CONTRACT RS90159001
NOTE 62p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Academic Achievement; Context Effect; Ethnic Groups; *Minority Groups; National Surveys; *Racial Differences; Scores; Secondary Education; *Social Influences; Student Characteristics; *Test Results; Test Use
IDENTIFIERS High School and Beyond (NCES); *National Assessment of Educational Progress; National Education Longitudinal Study 1988; Self Report Measures

ABSTRACT

This paper investigates the adequacy of the National Assessment of Educational Progress (NAEP) for taking into account dissimilarities in students' family, school, and community contexts when reporting test score differences among population groups (i.e., racial and ethnic minorities). This question was addressed by comparing the NAEP to other representative data for grades 8 and 12 from the National Education Longitudinal Survey (NELS) and High School and Beyond (HSB), studies that contain richer social context measures. These analyses show that the NAEP lacks a number of important social context measures and that the equality of some (but by no means all) of NAEP's measures is low because of reliance on student self-reports and other unreliable data sources. These weaknesses of the NAEP have important practical implications. Compared to HSB and NELS, the NAEP usually overestimates the achievement differences between students who come from different population groups but similar social contexts. However, at the secondary school level at which these analyses were conducted, these overestimates reflect primarily the NAEP's lack of important measures rather than its reliance on student self-reports. (Contains 12 figures, 2 tables, and 53 references.) (Author/SLD)

TM

TM

REPORTING MINORITY STUDENTS' TEST SCORES: HOW WELL CAN THE NAEP ACCOUNT FOR DIFFERENCES IN SOCIAL CONTEXT?

Mark Berends and Daniel Koretz
RAND

Forthcoming in *Educational Assessment*, 1996, 3(3)

The research reported here was undertaken by the Technical Review Panel (TRP) of the National Assessment of Educational Progress and was supported by the National Center for Education Statistics Contract No. RS90159001. The TRP is a joint endeavor of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles; RAND; and the University of Colorado at Boulder. The authors are grateful for the comments of Bob Calfee, Sharif Shakrani, Larry Ogle, Arnie Goldstein, Mary Naifeh, and the anonymous reviewers of *Educational Assessment*. We also appreciate the research assistance provided by Devah Pager, Stephanie Williamson, and Cathy Rosander and the programming by Elizabeth Lewis Harris. Requests for reprints should be sent to Mark Berends, RAND, 2100 M Street, NW, Washington, DC 20037-1270.

Abstract

This paper investigates the adequacy of the National Assessment of Educational Progress (NAEP) for taking into account dissimilarities in students' family, school, and community contexts when reporting test score differences among population groups (i.e., racial and ethnic minorities). This question was addressed by comparing the NAEP to other representative data for grades eight and twelve -- the National Educational Longitudinal Survey (NELS) and High School and Beyond (HSB) -- which contain richer social context measures. Our analyses show that NAEP lacks a number of important social context measures and that the quality of some (but by no means all) of NAEP's measures is low because of reliance on student self-reports and other unreliable data sources. These weaknesses of NAEP have important practical implications: compared to HSB and NELS, NAEP usually overestimates the achievement differences between students who come from different population groups but similar social contexts. However, at the secondary school level at which these analyses were conducted, these overestimates reflect primarily NAEP's lack of important measures rather than its reliance on student self-reports.

Reporting Minority Students' Test Scores:

How Well Can the NAEP Account for Differences in Social Context?

Concern about inequalities in educational outcomes among subgroups of the population has been heightened by the growing diversity of the United States population and the continuing dissatisfaction with the overall level of achievement of American students. Current reforms stress both raising the achievement of the entire population while reducing disparities among groups, which is certainly a daunting challenge (Smith and Scoll, 1995). In monitoring the academic progress of the nation's students, perhaps no other indicators have received more attention than standardized test scores in core subject areas such as mathematics and reading. While such overriding attention to test scores may be myopic if it lessens attention to other important indicators of educational success (Murnane, 1987), nationally representative achievement levels for all students and subgroups can be used "to foster a broader, more informed, and sustained discourse about the means and ends of education" (Bryk and Hermanson, 1993, p. 453). However, how should such group differences be reported?

Since its inception in 1969, the National Assessment of Educational Progress (NAEP) has been the only nationally representative, ongoing, and frequent assessment of the knowledge of American youth. One of the main functions of the National Assessment has been the reporting of test scores for various population groups (or "racial/ethnic" groups), such as African Americans, Hispanics, and whites.[1] Over time, the reporting of trends in achievement has become a central function of NAEP, and NAEP reports routinely present trends in the achievement of major population groups and discuss differences among them.

For certain purposes, however, reporting only unadjusted differences between population groups may be misleading because these groups tend to come from substantially different family, school, and community contexts, and these contextual differences are in turn powerful predictors of achievement. For example, many studies of academic achievement have found family characteristics to be strong correlates of minority and non-minority achievement levels (Coleman *et al.*, 1966; Sewell, Haller, and

[1] We avoid the conventional but misleading labeling of these groups as "racial or ethnic" and instead call them simply "population groups." The common categorization of individuals into these categories represent changing social conventions that have racial and ethnic components but are not well aligned with either racial or ethnic distinctions. For example, in the United States, it is conventional to identify people of mixed caucasion and negroid ancestry as black or African-American, nearly regardless of their relative proportions of caucasion and negroid background. South Asians who are racially caucasion are often labeled as Asian rather than white. South Americans with largely Native American (and hence racially Asian) background are often identified as white. "Hispanic" is typically considered an ethnic rather than a racial category, but in fact it encompasses groups with very different ethnic backgrounds, such as Cuban-Americans and Mexican-Americans. In these cases, the divergence from actual racial or ethnic classifications reflects currently dominant social conventions, some of which have varied over time. It is clearer as well as vastly simpler to retain the conventional groupings while discarding the misleading label of "racial/ethnic" than to attempt to use classifications that are more accurately racial or ethnic.

Portes, 1969; Sewell, Haller, and Ohlendorf, 1970; Jencks *et al.*, 1972, 1979; Averch *et al.*, 1972; Sewell and Hauser, 1975, 1980; Sewell, Hauser, and Wolf, 1980; Hauser, Tsai, and Sewell, 1983; Coleman, Hoffer, and Kilgore, 1982; Coleman and Hoffer, 1987). On average, African-American and Hispanic students have lower test scores than non-Hispanic whites, but they also tend to come from homes with parents who have less income and education and lower-status occupations -- and these parental characteristics are themselves powerful predictors of lower student achievement. Therefore, white and minority student test score differences that statistically adjust (or control) for dissimilarities in social context are typically far smaller than the unadjusted (raw) population-group differences.

Through the years, the National Assessment's own reports have typically presented unadjusted differences among population groups, without attempting to adjust them for dissimilarities in social context.[2] However, NAEP data can be used to adjust for differences in social context when reporting differences among population groups because NAEP has routinely gathered social context information in addition to test scores, and secondary analysts will no doubt use NAEP data this way even if NAEP's own reports do not. Available measures in the NAEP include students' family characteristics (e.g., parental education and single parent household), school and community measures (e.g., school composition, size and type of community, region), and school organization (e.g., curricular track and instructional experiences).

[2] A recent proposal to present state means adjusted for demographic differences generated substantial controversy and was in the end rejected by the National Assessment Governing Board.

Questions have been raised, however, about the adequacy of NAEP's social context measures for certain purposes. If social context is poorly measured, then the NAEP will not produce adequate estimates of the differences in scores among population groups holding these factors constant. Typically, differences among students from different population groups but similar social contexts would be overestimated.

In this study, we examine the adequacy of NAEP's measurement of social context for the specific purpose of reporting adjusted test-score differences among population groups from similar social contexts -- that is, score differences among these groups, holding constant differences in social context. This can be broken down into several separate questions:

- For this purpose, how adequate are NAEP's social context measures? That is, do they measure what they need to for this purpose?
- If there are problems involving the selection of social context constructs or their measurement, how much practical impact do they have? Specifically, in this case, do they substantially affect NAEP's estimates of performance differences among students from different population groups but similar social contexts?
- How accurate are the NAEP measures? That is, for present purposes, how do students' reports of family characteristics compare to those of their parents? Is there any practical impact of relying on student-reported information when adjusting the test score differences among population groups?

This study used several different methods to address these questions. The main body of analysis explored the practical impact of NAEP's choice of social-context constructs. This was accomplished by comparing analyses of the 1990 NAEP (grades 8 and 12) with analyses of the NELS 1988 eighth graders and the 1980 HSB base year senior cohort, both of which contain richer and stronger data on family and school characteristics than does NAEP. Specifically, reading and mathematics scores were separately regressed on population group membership (African-American, Hispanic, and non-Hispanic white) and a wide array of social-context variables to determine the extent to which holding social context constant shrank the unadjusted score differences between population groups. Because of their stronger social context information, HSB and NELS were treated as benchmarks to explore the extent to which NAEP could replicate results obtained with those databases. The regressions in all three databases relied on information provided by students and school administrators and excluded parental reports, so differences among the results from NAEP, HSB, and NELS primarily reflected the choice of social-context variables rather than inaccuracies in the source (e.g., student or principal reports). To put these contrasts into perspective, we also compared the unadjusted population-group differences in scores across the databases. Unsurprisingly, these raw differences sometimes differed among the databases, so regression results are reported in a form that takes those disparities into account.

Of course, differences among the results in the three data sets could stem from other factors as well, such as unintended sampling differences, differences in the ways in which constructs were operationalized in the three surveys (e.g., how questions were phrased and how responses were categorized), differences in the achievement tests included in the surveys, and the passage of time between the surveys (particularly in the case of grade 12). We could not systematically disentangle the effects of these factors, but there are several reasons to believe that the choice of variables was usually a primary source of differences in results. With one important exception (described in detail below), the results were quite similar across the numerous contrasts we drew. Many of the variables omitted from NAEP that appear to be important -- for example, family income -- would be expected to be significant on the basis of theory or other research. The sampling of the three surveys was in many respects similar and is generally considered to be of high quality. Finally, the operationalization of many of the variables included in our models was quite similar across the surveys. (Important exceptions to this generalization are noted below).

Several additional analyses explored the adequacy of student self-reports, upon which NAEP relies for most social context information. We reviewed extant studies of the consistency of information from students and parents in the National Educational Longitudinal Survey of 1988 (NELS) and the High School and Beyond (HSB) survey and complemented these with our own analyses of these data. To address the practical impact of reliance on student self-reports, we conducted parallel regression analyses of NELS and HSB using parent and student data, but these last analyses were seriously constrained by the limited number of relevant variables about which information was available from both sources.

DATA, MEASURES, AND METHODS

This section discusses the NAEP, HSB, and NELS samples; the operationalization of the social context measures analyzed across the databases; and our methodological approach.

SAMPLES

The National Assessment of Educational Progress

The 1990 NAEP, a nationally representative cross-sectional sample, assessed student performance in the subject areas of reading, mathematics, science, and writing of students at specific ages and grade levels in public and private schools (see Johnson and Allen, 1992). Students were selected from overlapping grade-eligible (grades 4, 8, and 12) and age-eligible samples (9, 13, and 17 years old).[3] For our purposes, we used the samples of students in the main national assessment who took either the mathematics test or the reading test in grades 8 or 12. These four samples ranged roughly from 6,300 to 6,500 students. Information about students' social contexts were obtained through a student background questionnaire administered at the time of the assessment and a school administrator survey.

The High School and Beyond Survey

In order to provide a benchmark to which we could compare the twelfth grade NAEP, we relied on the 1980 senior cohort in the High School and Beyond (HSB) survey. HSB is a two-stage stratified probability sample with schools as the first-stage units and students within schools as the second-stage units. The total sample comprised approximately 28,000 students in 1,100 schools. Some types of schools were over-sampled to ensure that adequate numbers in certain sub-populations. This paper

[3] NAEP also includes a "Trial State Assessment" (TSA), which consists of state-representative samples in limited number of grades and subjects in a majority of states. The TSA is not nationally representative, however, because of non-participating states, and it is not used in the present analyses.

used data from the base year senior cohort, collected in 1980 (for further description of the sample see Jones *et al.*, 1980).

The National Education Longitudinal Survey

The National Educational Longitudinal Survey base-year cohort of 1988 (NELS) provided a nationally representative database to which we could compare the eighth grade 1990 NAEP sample to assess its adequacy for portraying the test score differences among population groups. We analyzed the restricted-use version of NELS which surveyed and tested about 25,000 eighth grade students in 1,035 American schools; parent, teacher, and school surveys were also administered. Sponsored by the National Center for Education Statistics, NELS was a two-stage, stratified probability sample with schools selected as the first-stage unit and students within schools as the second-stage unit (for more details see Ingels *et al.*, 1990). After random selection of schools, about twenty-six eighth graders within each school were randomly selected; if schools had fewer than twenty-six students, all eligible students were included. Similar to HSB, some schools were over-sampled to ensure adequate samples of certain subgroups (e.g., Hispanics, Asians, and students attending private schools).

Parent Data

While NAEP has yet to gather data on a nationally representative sample of parents, NELS obtained information from over 90 percent of the students' parents (or other adult in the household) in its base year sample, and HSB contains information from a random subsample of students' parents. These data are useful for examining consistency between student and parent responses on measures of family characteristics. For the characteristics examined here, we assume that parents' reports will be more accurate those of students. For other factors, this may not be the case (Kaufman and Rasinski, 1991).

In both the NELS and HSB parent surveys, only one adult was administered the survey. In most cases, this was the parent who was most informed about the child's schooling. Most of the parent respondents were mothers (79% in NELS, 59% in HSB) or fathers (18% in NELS, 37% in HSB), but other female and male guardians were surveyed as well (3% in NELS, 4% in HSB). We created measures that were based on the respondent's reports for the mother and for the father. Therefore, statistics that report agreement between parent and child are based on reports about that parent by the one adult respondent, who might or might not be the adult who was the focus of the question. Information on consistency of reports among adults in the households is lacking.

VARIABLES

We used several sets of social-context variables in NAEP, HSB, and NELS, including measures of family background and composition, language use, community and school characteristics, and curricular placements. The variables within these sets are not identical across the three data sets; in particular, HSB and especially NELS contain richer sets of family and schooling variables than NAEP (see Table 1). In addition, in a few instances when a measure was available across databases, its operationalization differed.

Table 1

Test Scores -- Mathematics and Reading

The mathematics test in the 1990 NAEP aimed at assessing a variety of students' mathematical abilities such as understanding concepts, procedures, and problem solving (NAEP, 1988; Mullis *et al.*, 1991). NAEP reading performance was assessed in different domains such as reading a literary text, an informational document, and instructions to carry out a task (NAEP, 1989; Foertsch, 1992).

The 1980 HSB also tested students in the areas of mathematics and reading (see Heyns and Hilton, 1982). The mathematics test was in two parts lasting nineteen minutes. The first part contained twenty-five items measuring basic mathematical skills, asking questions primarily about which of two quantities is greater, or equal, or that the data given are insufficient to make a decision. The second seven-item section was comprised of more advanced mathematics questions. Requiring fifteen minutes to complete, the reading test consisted of twenty items that primarily measured student comprehension of short reading passages. The test scores provided by the National Center for Education Statistics (NCES) corrected for guessing.

NELS also tested students in various subject areas (see Rock and Pollack, 1991). The mathematics test lasted thirty minutes and contained forty items requiring students to make quantitative comparisons and to provide answers to word problems, diagrams, and calculations. The reading test lasted twenty-one minutes and consisted of twenty-one multiple choice items that measured student comprehension and interpretation of five short passages that varied in length from one paragraph to a half-page. Both mathematics and reading scores corrected for guessing.

Test scores can be placed on any number of scales, many of which are arbitrary and not directly comparable. Because the NAEP scaled-scores used in most NAEP reports are not comparable to NELS and HSB test scores, we standardized the scores in each subject area and database to have a mean of zero and a standard deviation of one. Thus, all scores have been put into a metric that can be directly compared and that permits comparison of regression estimates as well. An additional complication, however, is that the raw (unadjusted) group differences sometimes varied across databases. In those cases, changes expressed as fractions of a standard deviation would not be comparable. Therefore, we used the percent change in the (standardized) group differences as the metric for evaluating the effects of controlling for social context.

Population Group

Although there have been numerous conventions for classifying population groups, which are changing due to the increasing differentiation in the United States population, we used the coarse categories of African American, Hispanic, and (non-Hispanic) white. All others were grouped together in a residual "other" category that included Native Americans, Alaskan Natives, Asians, and Pacific Islanders. Dummy variables were created for each group. This four-category classification is an oversimplification. (For an examination of the educational outcome differences *within* these population groups, see Mare and Winship[1988].) However, this suffices for current purposes, and the databases are not sufficient to support a substantially more fine-grained classification.

Family Background

Measures of family background that were available in NAEP, HSB, and NELS included mother's and father's educational attainments, which were coded similarly with dummy variables.

Parents' occupations and family income were available in HSB and NELS, but not in NAEP. There were eighteen possible occupational categories for parents' occupations in HSB and twenty-one categories in NELS. In each database, dummy variables were created for each occupational category.

The measure of family income differed across HSB and NELS as well. In HSB, student-reported family income was measured using dummy variables for seven response categories ranging from less than $7000 (the reference) to $38,000 or more (1980 dollars). The parent measure in HSB differed from the student measure in that the parent questionnaire asked several questions about the sources of income (wages, tips, and salary; personally owned business and farm; dividends; interest; rents; alimony; and government aid). Each of these sources of income had nineteen response categories ranging from zero to $500,000 or more. We combined the various sources of parent-reported family income and collapsed the categories to seven to better match the student item even though the range of the categories remained dissimilar. In NELS, family income was only included in the parent survey. There were seventeen categories ranging from no income to $200,000 or more; these were dummy coded and translated into 1980 dollars.

Family Composition

All three databases included a measure for family composition -- whether the student lived in a two-parent household. HSB and NELS included an additional measure of the number of siblings.

Language Use

NAEP had the simplest measure of language use; NELS included a fairly complex array of measures; and HSB fell in between. The sole NAEP measure asked about the frequency of speaking a language other than English in the home; students could answer never, sometimes, or always. The HSB language use measures were based on two questionnaire items that asked what language the student usually spoke in the home and what language the student usually spoke at present. We created dummy variables for speakers of English (reference), Spanish, or some other language. In NELS, we relied on two items to measure language use, one asking about the language the student usually spoke in general and one about the language usually spoken at home, which provided more detailed response categories than HSB. Dummy variables were created for the following categories: usually spoke English (reference), usually spoke English and Spanish, usually spoke Spanish, or usually spoke some other language.

Community Characteristics

Community characteristics in NAEP, HSB, and NELS included region of the country and locale (i.e., urban, rural, suburban). NAEP includes an additional measure for size and type of community (STOC). The STOC composite relied on information about community size and type to create categones for areas of extreme rural, lower class metropolitan, upper-class metropolitan, urban fringe, main big city, medium city, and small place. Dummy variables were created for each category; upper-class metropolitan was the reference category.

The STOC variable in NAEP warrants special mention. As our results below indicate, STOC is a reasonably powerful predictor of achievement. Nonetheless, it has been the focus of intense debate, and its validity is now widely doubted. STOC is based in part on principals' estimates of the occupational profiles of their students' parents, and there is no evidence that principals have sufficient knowledge of that information. In addition, STOC categorizes communities inconsistently over time and inconsistently with other data, which makes substantive conclusions based on the STOC variable problematic (for discussion, see Berends, Koretz, and Lewis, 1994; Koretz, 1991; Lippman, 1993). We include STOC in some of our models, but because of its questionable validity, we note instances in which the results appear especially influenced by its inclusion.

School Characteristics

School characteristics common to NAEP, HSB, and NELS included percent of students within the school receiving free or reduced-price lunch, school sector (private versus public), demographic composition (percent black and Hispanic), and school mean achievement.

HSB and NELS, but not NAEP, had additional measures for aggregated school socioeconomic status and a school attendance rate. The aggregate SES measure was the within-school mean of the student-reported SES composite (an unweighted standardized composite based on student-reported mother's and father's education levels, father's occupation recoded into a Duncan SEI scale [Duncan, 1961], family income, and a scale of household-possession items).[4]

Curricular Differentiation

Research has shown that high school tracking is related to different student experiences both academic and social (for a review see Gamoran and Berends, 1987; Oakes, Gamoran, and Page, 1992). Studies of HSB have shown these schooling experiences mediate the effects of family background and the characteristics of schools and community (Gamoran, 1987; Lee and Bryk, 1988) and partially explain the differences in outcomes among population groups (Oakes, 1990; Gamoran and Mare, 1989).

In NAEP (12th grade only) and HSB, students reported whether they were an academic, general, or vocational curricular program. Dummy variables were included for each curricular category. Additional measures available across the three databases included the percentages of students in talented and gifted classes, remedial reading classes, or remedial math classes. The eighth grade NELS data contained richer data on students' placements in ability groups in both mathematics and reading. Dummy variables were created for student reports of grouping in both subjects; categories were high, middle, or low group, not grouped, and don't know. Further variables were created from students yes-no responses for participation in an advanced mathematics or English class or remedial mathematics or English class.

[4] In the regressions the school measures for SES and achievement were adjusted school means that subtracted the individual respondent from the pooled mean so that any individual was not part of the school mean when predicting his or her test score.

METHODS

Regression analyses of all three databases were used to estimate the changes in the mean test score differences between whites and both African-Americans and Hispanics when controlling for various sets of social context measures. Separate analyses were conducted for mathematics and reading scores. We estimated a series of models that cumulatively added an increasingly wide array of social context variables. Each stage added part or all of a set of variables (e.g., the set of family background variables). First, to provide a baseline to compare the adjusted estimates, the unadjusted mean test score differences between population groups were calculated. Second, certain family background characteristics were included in the regression, such as parents' education. A further model added family composition measures (two-parent household and siblings) and language use. Subsequent models cumulatively added school and community characteristics, school mean achievement (using the mean only for the subject area in question), and curricular differentiation measures. At each stage, we estimated the reduction in the unadjusted (raw) difference between whites and either African-Americans or Hispanics.

Comparisons of these regression models across the three databases provided an indication of the adequacy of the NAEP variable sets. To the extent feasible, the models were similar across the three databases. NELS and HSB regressions served as baselines, and we attempted to replicate them using NAEP. The more limited sets of variables in NAEP, however, necessarily gave us less fully specified models at some stages, and at those stages, we predicted that the NAEP regression would reduce the raw differences between groups correspondingly less than the corresponding NELS and HSB regressions.

We pursued two different approaches in evaluating the impact of NAEP's reliance on student self-reports of social context variables. First, we directly compared student and parental reports of the same variables. Because NAEP did not survey parents, we had to use NELS and HSB data for this purpose. (Even apart from its lack of a parent survey, NAEP generally has little overlap across data sources, which restricts the analysis of NAEP data quality that can be done using only the NAEP data itself.) We reviewed comparisons of parent and student responses in NELS (Kaufman and Rasinski, 1991) and HSB (Fetters, Stowe, and Owings, 1984). In addition, because our operationalization of certain measures differed from these previous studies, we also conducted our own analyses comparing student and parent reports.

Second, we assessed the impact of using student-reported data by conducting parallel regression analyses, one set using student reports and the other using parent reports. The disparity between the two sets of regressions was a measure of the practical importance, *for these specific purposes*, of NAEP's reliance on student reports.

These two types of analyses, however, can only begin to answer our questions about the data quality of individual social context measures because our comparisons between students and parents were limited to a subset of the measures. More complete data on other social context measures from multiple sources are needed to fully address this issue.

RESULTS

In what follows, first the results presented are the comparisons of regression analyses in NAEP, HSB, and NELS directly evaluate the relative adequacy of NAEP for reporting differences among population groups after controlling for a wide variety of social context variables. The second set of results addresses the quality of the student self-reports upon which NAEP relies by comparing students' responses with those of their parents. A final section assesses the consequences of relying on student- versus parent-reported family characteristics when estimating the adjusted test score differences between population groups by comparing regressions that use student-reported and parent-reported data.

TEST SCORE DIFFERENCES:
THE IMPACT OF CONTROLLING FOR SOCIAL CONTEXT

This section presents analyses bearing on our core question: how adequate is NAEP, compared to NELS and HSB, for taking social context into account when presenting test scores? We examined the impact of controlling for family, school, and community characteristics on black-white and Hispanic-white achievement differences in grades eight (NAEP and NELS) and twelve (NAEP and HSB). As noted earlier, differences are expressed as fractions of the (pooled) standard deviations, and the impact of controlling for social context is expressed as the percentage reduction of the unadjusted differences between groups.

The unadjusted differences in mathematics and reading between groups varied appreciably among the data sets. The unadjusted differences between African Americans and whites were largest in the eighth grade: from .77 to .93 of a standard deviation (SD) in mathematics and from .44 to .67 of a SD in reading. The unadjusted differences between Hispanics and whites varied as well. The Hispanic-white difference varied most in twelfth-grade reading: .51 SD in NAEP but .73 SD in HSB.

The effects of controlling for family, school, and community characteristics also varied somewhat among the three databases. In all three data sets, controlling for these social context measures substantially reduced the test score differences between groups, but the extent of the reduction varied. Our comparisons indicate that in most instances, NAEP provides a smaller and less adequate adjustment of the differences between groups and suggest that the reason may be NAEP's lack of measures of such social context factors as parent occupation, income, number of siblings, and measures of ability grouping arrangements in the school.

The following sections report three findings: the unadjusted differences between groups (in proportions of a standard deviation [SD]); a series of adjusted differences, cumulatively controlling for an increasing number of social context variables (also in proportions of a SD); and the percentage reduction in the unadjusted group difference brought about by controlling for these variables. Black-white differences in eighth grade mathematics and reading (using NAEP and NELS) are presented first and in the greatest detail. The following sections present the remaining contrasts (black-white differences in twelfth grade and Hispanic-white differences in both grades) in less detail, focusing primarily on mathematics for simplicity. (Results in reading were generally similar.)[5]

[5] Analyses of Asian-white difference in NELS are not reported here but are available from the first author upon request. The sample of Asians in NELS was large enough to be analyzed separately because of oversampling.

Black - White Test Score Differences in Grade Eight (NELS and NAEP)

The impact of controlling for differences in social context was considerably greater in NELS than in NAEP, largely because of variables included in NELS but not NAEP. Using the available information about social context reduced the unadjusted difference in scores by 10 to 15 percentage points more in NELS than in NAEP, depending on the measures included in the regression.

As noted earlier, many of the unadjusted differences between groups varied from one database to another, and that was the case in eighth-grade mathematics: The black-white differences were .93 of a SD in NAEP and .77 of a SD in NELS. This is shown in the left-most pair of points in Figure 1 (above the label "unadjusted difference"). In the original metrics, these differences translate into about thirty points on the NAEP mathematics test and just under seven points on the NELS mathematics test.

Adjusting these differences with the available measures of family background reduced them substantially, although more in NELS than in NAEP. The gap was reduced to .78 in the NAEP and .52 in NELS (shown in the second set of points in Figure 1, over the label "adjusting for family background"). The proportionately greater reduction in NELS (evident in the steeper downward slope of that segment of the NELS line in Figure 1) is quantified in the corresponding bars in Figure 2: the NAEP family background model resulted in a 16 percent reduction of the unadjusted difference, while the NELS model reduced it by 32 percent. The fact that NELS contains information on parents' occupations and family in..me accounts for the greater reduction in NELS.[6]

Adding family composition (two-parent household and number of siblings) and language use in the home to the family background measures decreased the unadjusted achievement difference by only a few percentage points more. This is evident in both the nearly flat lines between the second and third set of points in Figure 1 and the bars representing the cumulative percent reduction for "adjusting for family background" and "adding family composition and language use" that are practically indistinguishable from each other in Figure 2.

In contrast, adding controls for school and community characteristics substantially reduced the estimated gap between blacks and whites (Figure 1) -- in NELS, resulting in nearly a 50 percent cumulative reduction in the group difference (Figure 2). Adding school mean achievement had little impact, but adding curricular differentiation produced an appreciable further reduction in the group differences. The cumulative impact of controlling for all of these variables in NELS was a 56 percent reduction in the black-white difference (Figure 2): from .77 SD to .34 SD (Figure 1). In the original metric of the NELS test, the test score difference between blacks and whites was reduced from roughly seven points to three points. The corresponding reduction in NAEP, while large, was considerably smaller: 44 percent (Figure 2) -- from . 93 to .52 SD (Figure 1), or in the original metric from about thirty points to about seventeen points.

[6] Income alone reduced the unadjusted difference in scores by 10 percent.

This is one instance in which the reduction of the black-white difference in NAEP would have been considerably smaller if the questionable size-and-type-of-community (STOC) variable had been excluded from the analysis. Thus, without STOC, the adjustments yielded by the NAEP model would have fallen further short of those produced by NELS.

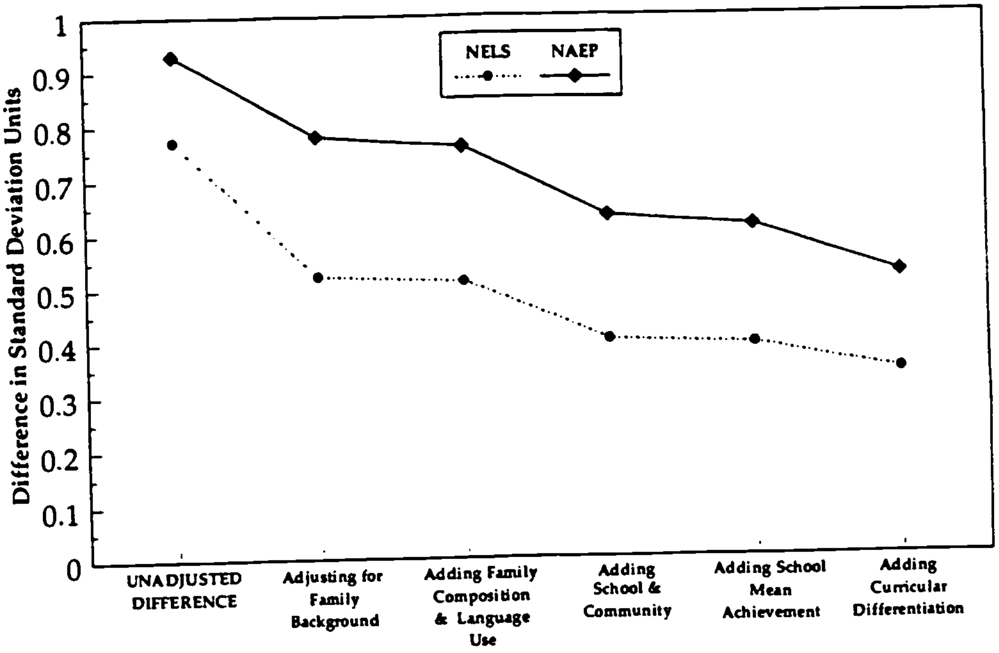

Figure 1 -- Mathematics Differences Between African Americans and Whites Unadjusted and Adjusted, Grade 8 (Differences Expressed in Standard Deviation Units)

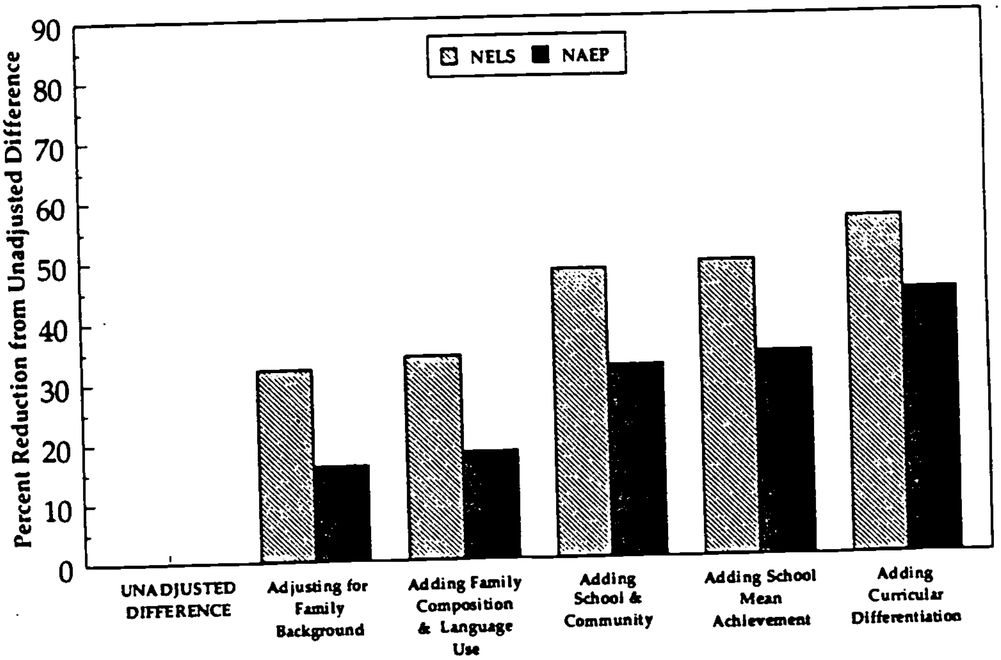

Figure 2 -- Cumulative Percent Reduction in Unadjusted Mathematics Differences Between African Americans and Whites, Grade 8

Black-white differences in eighth-grade reading differ from these mathematics results in important specifics but show some of the same general patterns. In contrast to mathematics, the unadjusted reading differences between blacks and whites are smaller, and the black-white gap in NELS is larger than that in NAEP. But in reading as in mathematics, adjusting only for differences in family background was sufficient to bring about a marked reduction in the gap, and here again, the reduction was proportionately much larger in NELS than in NAEP (Figures 3 and 4). Also similar to mathematics, the cumulative effect of controlling for all of the social context variables was considerably larger in NELS than in NAEP. Indeed, in proportional terms, the disparity between the databases in this respect was even larger in reading. The cumulative reduction in NELS was 54 percent (from .67 to .31 SD), while in NAEP it was 36 percent (from .44 to .28 SD; Figures 3 and 4). (In terms of the original metric, the NELS unadjusted reading difference of about three points was reduced to about one-and-a-half points; the unadjusted NAEP difference of sixteen points was reduced to ten points.)

In sum, these eighth grade analyses revealed that the measures of students' social contexts in NELS were richer than those in NAEP and reduced the test score differences between blacks and whites by a substantially greater percentage. Our comparisons suggest that information presently absent from the NAEP -- such as family income, parental occupation, and ability grouping measures -- contribute to the lesser reduction in the test score differences in NAEP.

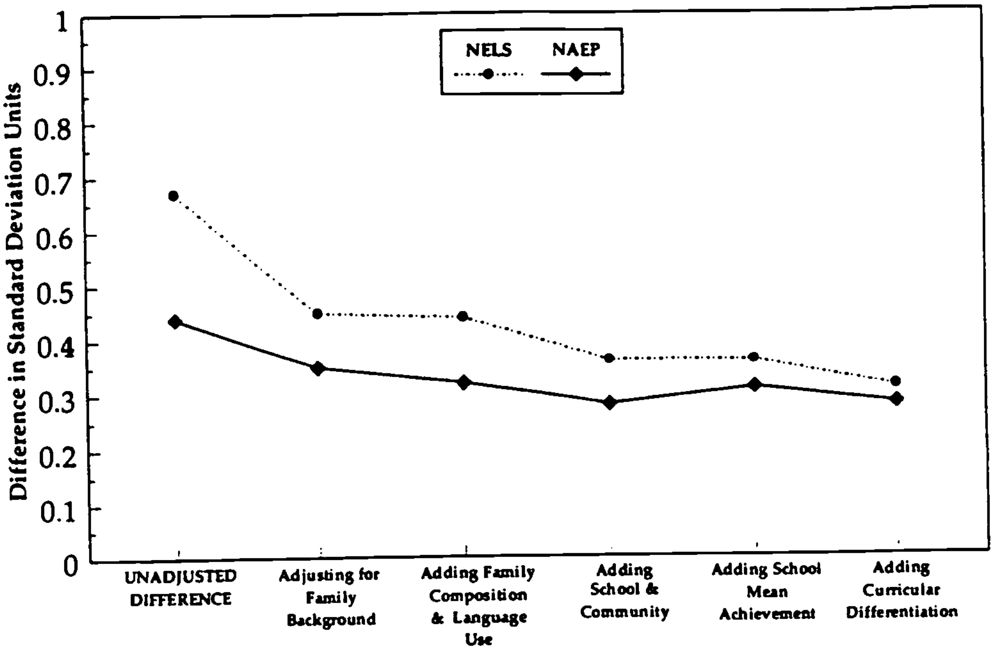

Figure 3 -- Reading Differences Between African Americans and Whites Unadjusted and Adjusted, Grade 8 (Differences Expressed in Standard Deviation Units)

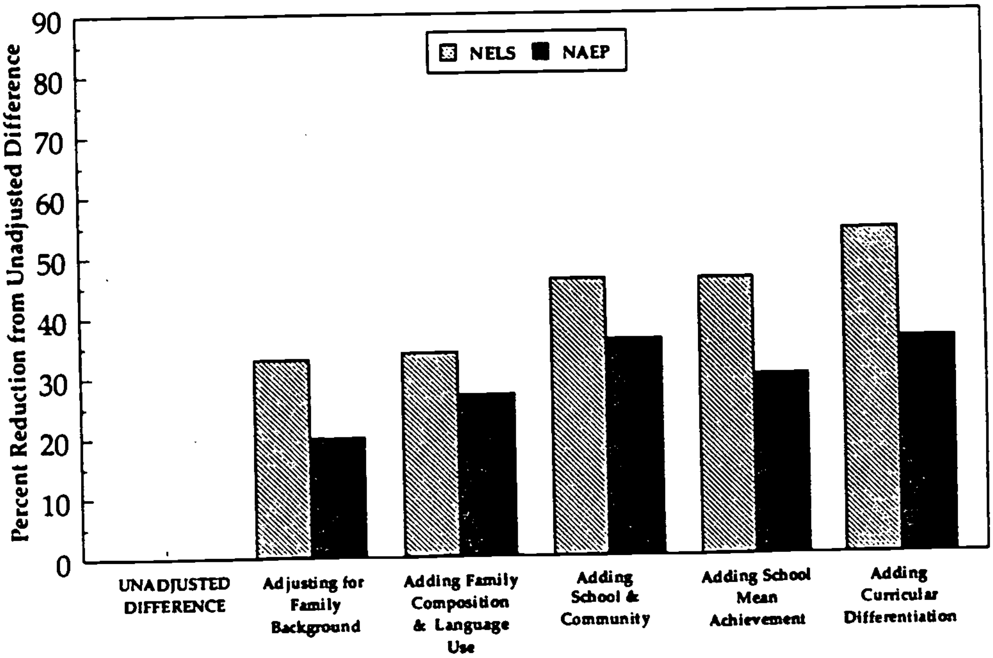

Figure 4 -- Cumulative Percent Reduction in Unadjusted Reading Differences Between African Americans and Whites, Grade 8

BEST COPY AVAILABLE

Black - White Test Score Differences in Grade Twelve (HSB and NAEP)

The unadjusted difference in mathematics between twelfth-grade African-Americans and whites was nearly identical in the HSB and NAEP: .83 SD in HSB (seven points in the original metric) and .86 SD in NAEP (twenty-nine points in the original metric) (Figure 5). In reading, the disparity between the databases was smaller than in the eighth grade: .78 SD in HSB (or eight points in original metric) and .66 SD in NAEP (or twenty-three points) (Figure 7). However, a variety of data indicate a decrease in the gap between blacks and whites during the fielding of HSB in 1980 and the 1990 iteration of NAEP (e.g., Rasinski, Ingels, Rock, and Pollack, 1993; Koretz, 1992; Mullis, Dossey, Foertsch, Jones, and Gentile, 1991). Therefore, it is likely that if less time had elapsed between the two surveys, the gap between the groups would have been relatively larger in NAEP, making the disparity between the databases larger in mathematics but smaller in reading.

Controlling for family, school, and community characteristics reduced the unadjusted differences in mathematics and reading scores by roughly 40 percent in both NAEP and HSB (see third set of bars in Figure 6). This is an instance, however, in which NAEP's reliance on the size-and-type-of-community (STOC) variable is particularly important; without that questionable variable, NAEP would not have controlled as adequately as HSB for social-context differences.

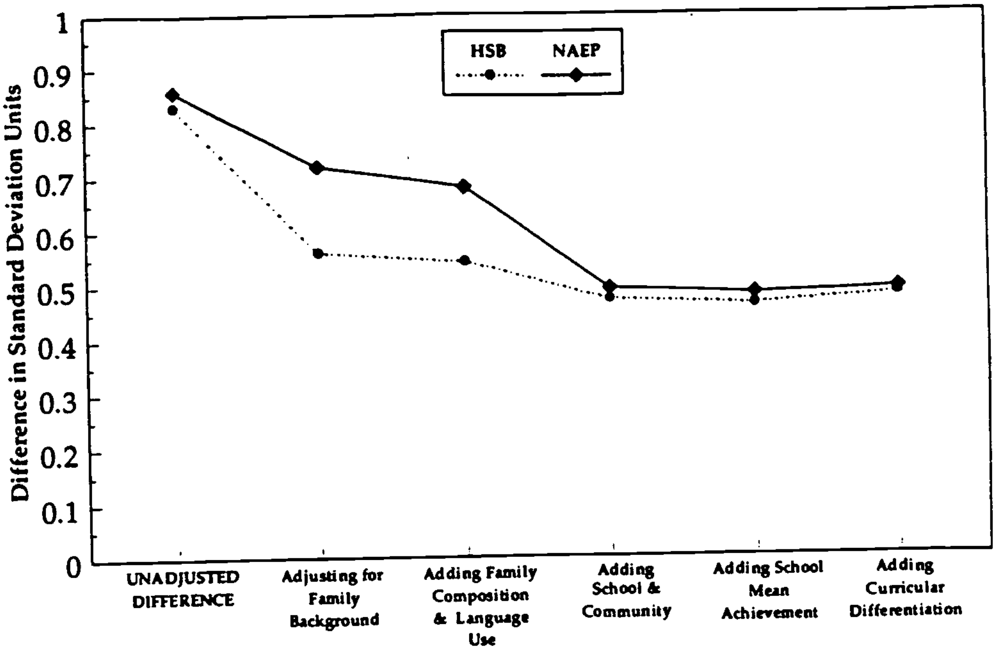

Figure 5 -- Mathematics Differences Between African Americans and Whites Unadjusted and Adjusted, Grade 12 (Differences Expressed in Standard Deviation Units)

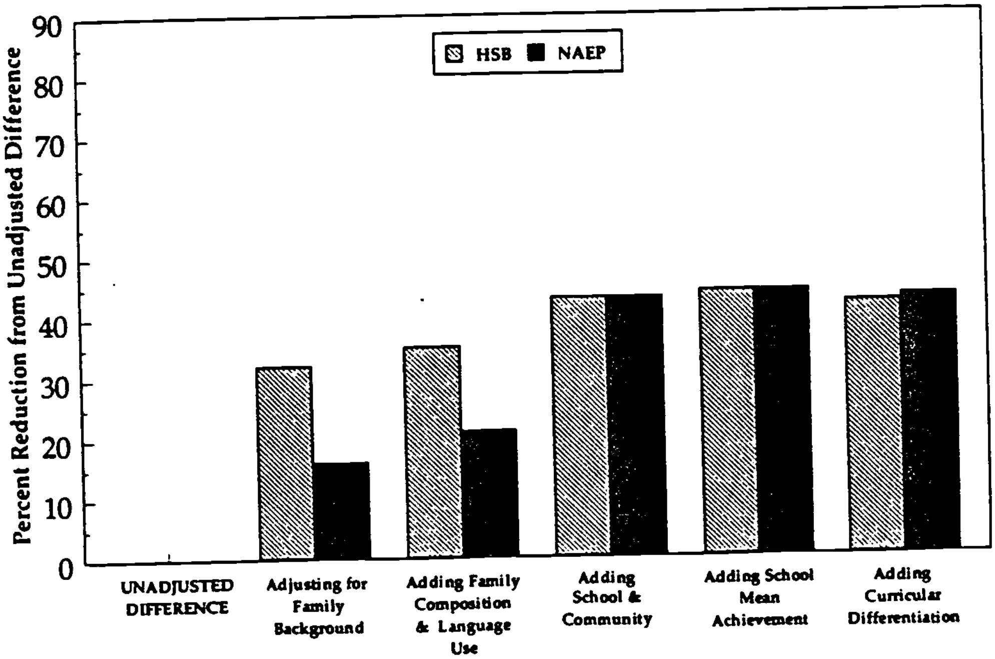

Figure 6 -- Cumulative Percent Reduction in Unadjusted Mathematics Differences Between African Americans and Whites, Grade 12

BEST COPY AVAILABLE

In mathematics, taking into account only the first set of social context variables -- family background -- reduced the black-white difference much more in HSB than in the NAEP (Figure 5). In HSB, adding only family background measures the black-white gap in mathematics from .83 to .56 of a SD (Figure 5). This corresponds to a 32 percent decrease (Figure 6). In contrast, holding constant differences in family background reduced the black-white mathematics gap in the NAEP by half as much (Figure 6). The larger reductions in HSB appear to reflect its inclusion of measures of parental occupation and family income, both of which are lacking in NAEP.

Adding controls for the second set of variables -- family composition and language use -- had similarly small effects in both NAEP and HSB (Figure 5). Neither HSB's more elaborate language variable nor its inclusion of a measure of the number of siblings was of any real consequence in this particular instance.

However, adding controls for community and school measures had *more* of an effect in NAEP than in HSB, as shown by the steeper slope for that part of the NAEP line in Figure 5. As a result, when all of the first three sets of variables (including community and school measures) were held constant, the total reduction in the black-white difference in mathematics was identical in the two databases: 43 percent (Figure 6).

A similar pattern appeared in reading. Here again, adjusting only for family background had a larger impact in HSB. (Note the steeper decline in Figure 7; see also the percent reductions in Figure 8). The black-white difference in reading, larger in HSB when no social context variables are considered (.78 SD, versus .66 SD in NAEP), was brought down to the level of NAEP's difference when family background was controlled. The effect of adding controls for family composition and language use was

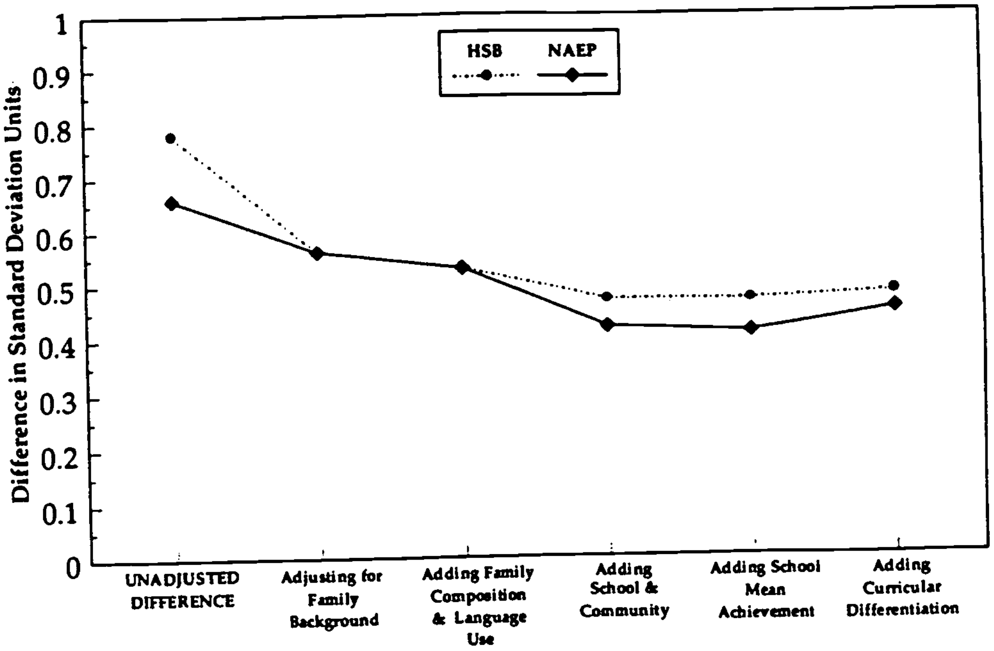

Figure 7 -- Reading Differences Between African Americans and Whites
Unadjusted and Adjusted, Grade 12
(Differences Expressed in Standard Deviation Units)

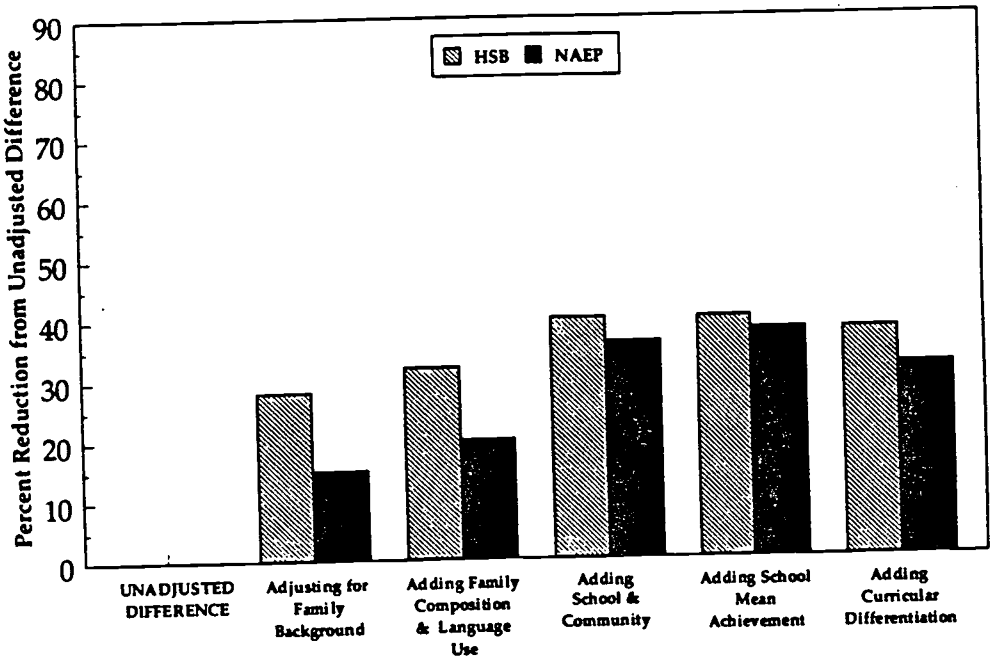

Figure 8 -- Cumulative Percent Reduction in Unadjusted Reading Differences
Between African Americans and Whites, Grade 12

BEST COPY AVAILABLE

identical in HSB and NAEP and again small. The impact of adding school and community variables was also larger in NAEP, although in this case the disparity between the databases was smaller. The cumulative reduction in the black-white difference after taking the first three sets of variables into account was similar as well: 36 percent in NAEP and 40 percent in HSB (Figure 8).

However, the specific measures for the school and community contexts were different in the two data sets. Measures such as region of the country, locale, school sector, and school socio-demographic composition were important in the HSB data for reducing the achievement gaps between population groups. The NAEP model included region, school population-group composition, school sector, percentage of students who were in a free lunch program, and size and type of community (STOC). Similar to the grade eight comparisons, the reduction of the black-white differences in the NAEP models would have been appreciably smaller (and thus smaller than the reduction in HSB) if STOC had been excluded.

Adding controls for additional social context variables (mean achievement and curricular differentiation had no appreciable effect on the estimated black-white difference in either HSB or NAEP.[7]

[7] The black-white gap *increased* slightly when adding curricular differentiation to the other variables already considered in HSB and NAEP. We explored this instance further in HSB, where we found significant interactions between track and the black/white dummy variable. African Americans in the academic track, where they are underrepresented, scored lower than their white counterparts in math. There were no significant interactions between track and population group in reading.

Hispanic - White Test Score Differences in Grade Eight (NELS and NAEP)

In eighth grade, the unadjusted mathematics differences between Hispanics and whites differed modestly between the two data sets (.70 SD in NAEP and .60 SD in NELS; Figure 9), but the unadjusted reading differences were practically the same (.53 SD in NAEP and .56 SD in NELS; Figure 10). Controlling for differences in social context reduced the mathematics and reading differences much more in NELS than in NAEP. This again suggests that several important social context measures are absent in the NAEP.

In mathematics, the unadjusted Hispanic-white difference in NELS is modestly smaller than that in NAEP, but the effect of controlling for social-context differences progressively widened the gap between the databases (Figure 9). After controlling for all five sets of social-text variables, the Hispanic-white difference in NELS was well under half that in NAEP: .16 SD in NELS compared to .38 SD in NAEP. The greatest reduction in NELS (after including all variables) was 73 percent; the greatest reduction in NAEP (after including school and community variables) was only 45 percent.[8] The small size of the final Hispanic-white difference in NELS is particularly striking.

The same basic pattern appeared in reading. The unadjusted Hispanic-white difference was nearly identical in the two databases (roughly .55 SD), but taking social context variables into account shrank the difference far more in NAEP, causing the two databases to diverge (Figure 10). The maximum reduction in the Hispanic-white gap

[8] For simplicity, we describe the percent reduction in this section of the paper but do not present bar graphs.

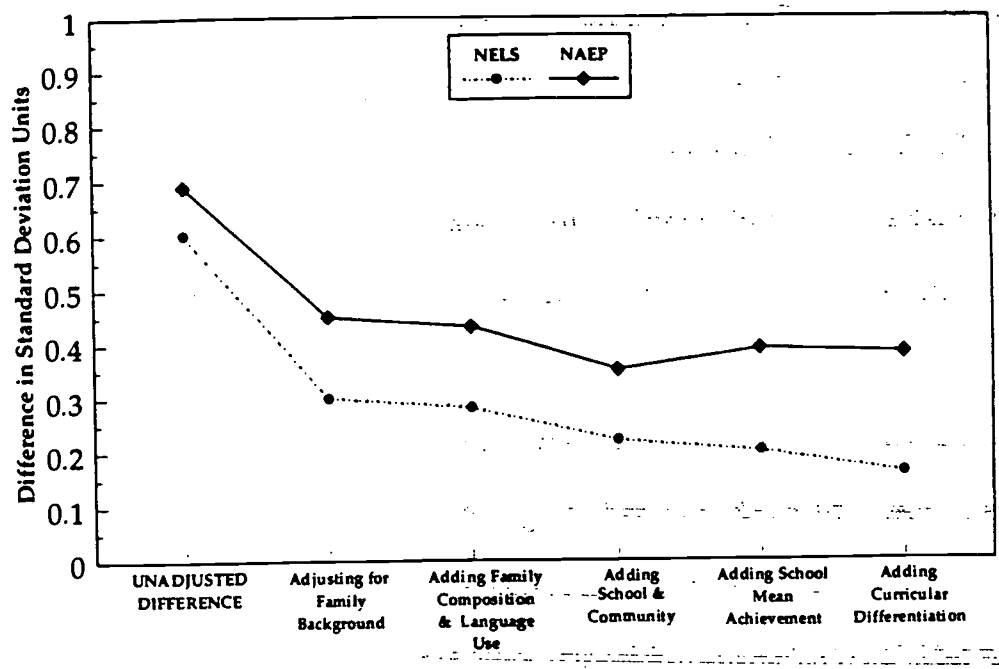

Figure 9 – Mathematics Differences Between Hispanics and Whites Unadjusted and Adjusted, Grade 8 (Differences Expressed in Standard Deviation Units)

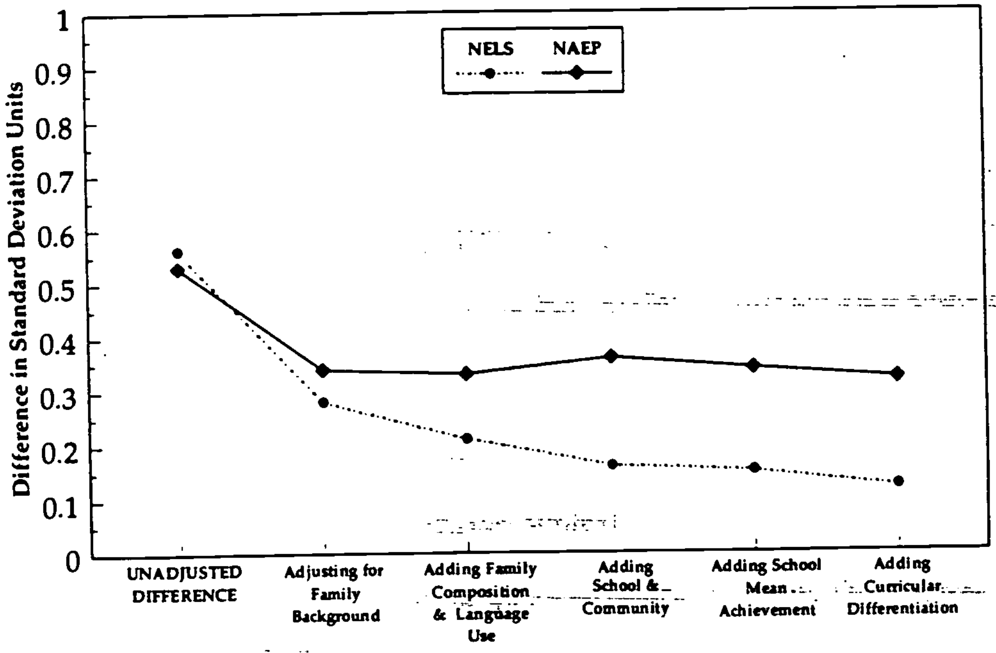

Figure 10 – Reading Differences Between Hispanics and Whites Unadjusted and Adjusted, Grade 8 (Differences Expressed in Standard Deviation Units)

BEST COPY AVAILABLE

was much larger in NELS: 79 percent, versus 40 percent in NAEP. And again, the final adjusted Hispanic-white difference in NELS was very small – only .12 SD.

The specific similarities and differences between mathematics and reading are also worthy of note. In both subjects, adjusting for the first set of variables (family background) had the greatest impact, and the effect was only modestly greater in NELS than in NAEP. Beyond that point, however, the mathematics and reading results differed substantially. In mathematics, language use had little impact; the additional divergence between NELS and NAEP stemmed from the last two sets in the models (school mean achievement and curricular differentiation; see Figure 9). In reading, by contrast, these last two sets of variables had little effect. The additional divergence in reading stemmed from the second and third sets: family composition and language use, and school and community characteristics (see Figure 10). Some of these differences are difficult to interpret, but is reasonable that adding the more elaborate language variables in NELS (variables for whether the student usually spoke Spanish, Spanish and English, English only, or English and some other language besides Spanish) helped reduce the gap between Hispanics and whites in reading.

In short, these comparisons show that the lack of measures of parent occupation, income, and ability grouping in the NAEP, as well as its inadequate language variables, seriously hindered its ability to portray Hispanic-white differences independent of social context.

Hispanic - White Test Score Differences in Grade Twelve (HSB and NAEP)

The twelfth-grade Hispanic-white contrast was the only case in which controlling for social context reduced the unadjusted test score differences *more* in NAEP than in the second database (in this case, HSB). In both mathematics (Figure 11) and reading (Figure 12), the nearly parallel lines between the points depicting the unadjusted differences and those adjusting for family background mean that the percent reduction in the group differences were similar. There was some divergence when adding measures of family composition and language use. The cumulative addition of the school and community measures made the lines diverge even more in mathematics and reading, but as we noted previously in the other comparisons, this was primarily due to the size-and-type-community measure in NAEP. The greatest reduction in the NAEP mathematics difference was from .60 to .20 SD in the school and community model (a 67 percent reduction), whereas HSB was reduced in this model from .65 to .35 SD (a 46 percent reduction). The pattern in reading was similar, where the greatest reduction in NAEP was also in the school and community model (from .51 to .22, a 57 percent reduction). However, the unadjusted difference in the HSB school and community model was reduced from .73 to .46 SD (a 37 percent reduction).

It is difficult to determine exactly why there was a greater reduction in the Hispanic-white achievement differences in the NAEP compared to HSB. One plausible reason is that this exception may stem from the increasing growth and changing composition of the Hispanic population between 1980 (HSB) and 1990 (NAEP) (Hernandez, 1993; McArthur, 1993; McDonnell and Hill, 1993; National Center for Education Statistics, 1993). However, this is only speculation that needs further analysis.

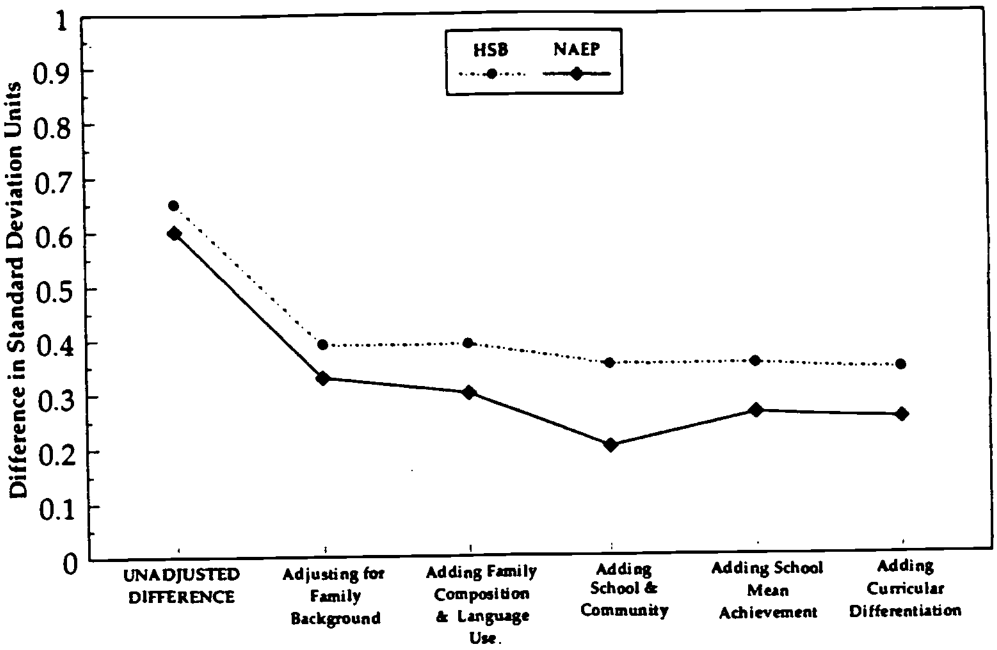

Figure 11 -- Mathematics Differences Between Hispanics and Whites
Unadjusted and Adjusted, Grade 12
(Differences Expressed in Standard Deviation Units)

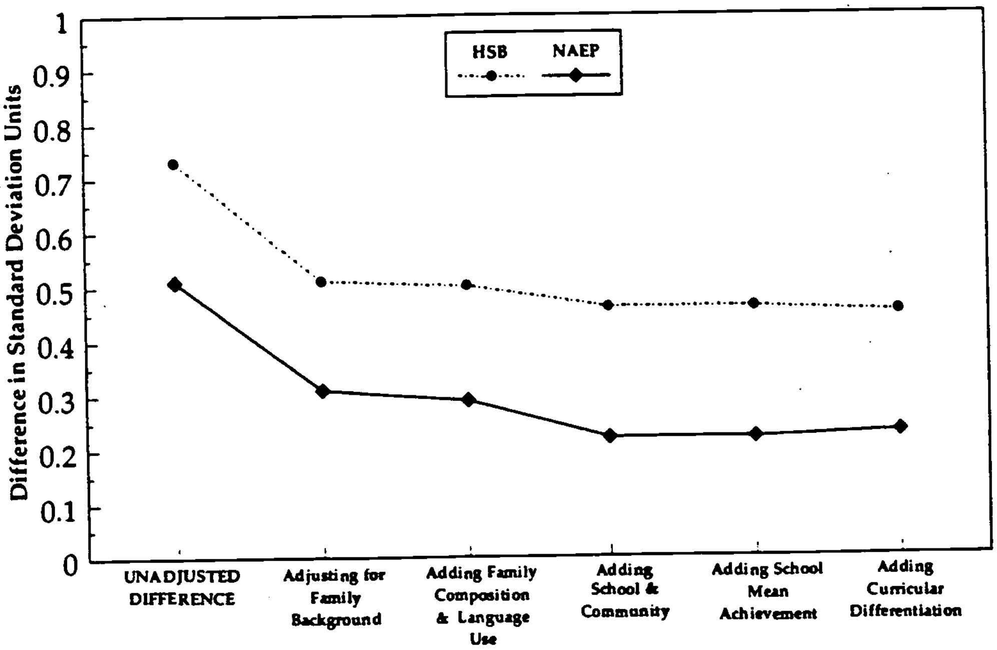

Figure 12 -- Reading Differences Between Hispanics and Whites
Unadjusted and Adjusted, Grade 12
(Differences Expressed in Standard Deviation Units)

THE QUALITY OF FAMILY CONTEXT MEASURES

In addition to these core analyses that explored the practical impact of NAEP's sets of social context measures on the population group test score differences, we also examined the impact of NAEP's reliance on student reports of family characteristics. This section reports the degree of consistency between students and parents on a limited set of measures, and the next section shows the practical impact of these differences in parallel regressions based on student and parent data.

The extent of agreement between students' and parents' responses varied substantially depending on the specific information sought. The relative consistency of students' and parents' responses was not closely aligned with the categorizations of variables used in our regression analyses. Rather, consistency was higher, as one might expect, when items asked for relatively apparent and discrete information. Students and parents were generally consistent about more obvious characteristics, such as population group membership, family composition (i.e., the number of parents and siblings in the household), and language use. Students and parents were less consistent with respect to family background characteristics (e.g., parents' education levels and occupations and family income). These findings, detailed below, are consistent with other studies that examined the quality of student reports of family characteristics and attitudes toward school in other databases (Cohen and Orum, 1972; Borus and Nestel, 1973; Kayser and Summers, 1973; Kerckhoff, Mason, and Poss, 1973; Looker, 1989).

The general agreement between students and parents about their population-group membership appeared in both grade levels (see Table 2). In NELS and HSB, parent and student responses to items about population-group membership matched exactly in over 90 percent of all cases (Fetters *et al.*, 1984; Kaufman and Rasinski, 1991).

Table 2

The consistency of parent and child responses about family background measures, however, was moderate to low and differed dramatically by grade level. Our analyses of NELS showed that responding parents and eighth-grade students agreed in their statements about mother's educational attainment 54 percent of the time. Consistency in parent-child responses increased with age; in HSB, 72 percent of the responses by seniors matched those of their parents. If we considered as matches student and parent's responses differing by one category (e.g. the parent reports some college and the student reports high school graduation), agreement rates increased to 80 percent in NELS and 96 percent in HSB (not shown). In NELS, the consistency of responses for father's educational attainment was slightly lower than for mother's education. For example, while eighth-graders agreed with parents about mothers' attainment 54 percent of the time, they agreed with respect to fathers' attainment 49 percent of the time.

Students' reports of income appeared to be even less consistent with parents' responses, but the information on this point was limited to the HSB and was ambiguous. (NELS gathered income data only from parents.) The HSB questions about income administered to students considered fewer sources of income than those given to parents and used different income categories. Although we collapsed the parent variable into categories that more nearly matched those in the student question, the resulting variables were still not fully comparable, and the differences between them undoubtedly eroded agreement between parents and students. The percent agreement between the collapsed parent and student responses was 31 percent. Counting as matches parents and students who differed by one income category, the matched responses increased to 60 percent.

Both eighth- and twelfth-grade students were inconsistent with their parents in their reports of parents' occupational status. In NELS, fewer than 45 percent of eighth graders agreed with the responding parents about mother's or father's occupation. The occupational items in NELS and HSB included nearly twenty possible response categories, however, and this level of detail contributed to the disagreement between parents and students. When some similar occupational categories were combined, resulting in a total of twelve categories, the percent agreement increased to about 60 percent (59 percent for mother's occupation and 58 percent for father's occupation).[9] Here again, consistency increased with age, but in this case the improvement was modest. For example, 52 percent of seniors--in contrast to the 45 percent of eighth graders noted above--agreed with their parents' report of mother's occupation.

Parent and student responses to family composition items (i.e., number of siblings and two-parent household) were quite consistent. In both NELS and HSB, 90 percent or more of the student responses about the number of parents living in the student's household agreed with those of their parents. Parents and students usually

[9] These estimates resulted from the following combinations: The laborer, operator, and craftsman categories were combined due to their similar occupational attributes. The two categories for professional occupation were also combined. If students reported that their mothers were homemakers, and mothers reported that they were clerical workers, teachers, service workers, or never worked, we considered these responses as agreeing.

also agreed about the number of siblings the student had (80 percent in NELS and 82 percent in HSB).

Agreement between parent and student responses to items about language use were quite high in all the data sets, although the responses were more consistent in HSB than in NELS. In HSB, virtually all students and parents (98 percent) agreed about what language was usually spoken in the home. In NELS, parents and students agreed 89 percent of the time on whether another language was usually spoken and 87 percent on the particular language usually spoken in the home. It is not clear whether the lower agreement in NELS reflects the younger age of the respondents or the specific questions employed.

THE PRACTICAL IMPACT OF RELYING ON STUDENT-REPORTED FAMILY CHARACTERISTICS

In addition to examining the consistency of parents' and students' reports of social context, we directly examined the practical impact of relying on student-reported rather than parent-reported measures of social context. We investigated this by comparing otherwise identical regressions based on information from parents and students. The regressions were similar to those above, in that they examined the impact of controlling for social context on the size of test-score differences between population groups, but they were necessarily limited to the subset of social context variables for which information was available from both parents and students. Because NAEP lacks parent reports, these analyses were conducted using NELS and HSB.

These analyses indicated that *for some specific purposes*, reliance on student-reported data rather than parent reports has very little effect. However, because of the limits of these analyses, these findings do not indicate that student reports are *generally* sufficient.

At both the eighth and twelfth grades, there were virtually no differences between regressions using student and parent reports. Figure 13 shows one case: the unadjusted and adjusted mathematics score differences between eighth-grade African Americans and whites based on student and parent sources. This is an illustrative case not only because it is similar to the main results reported above, but also because younger students and African American students have been shown by others to be less reliable than Hispanics and whites when reporting family background characteristics (Kaufman and Rasinski, 1991; Looker, 1989).

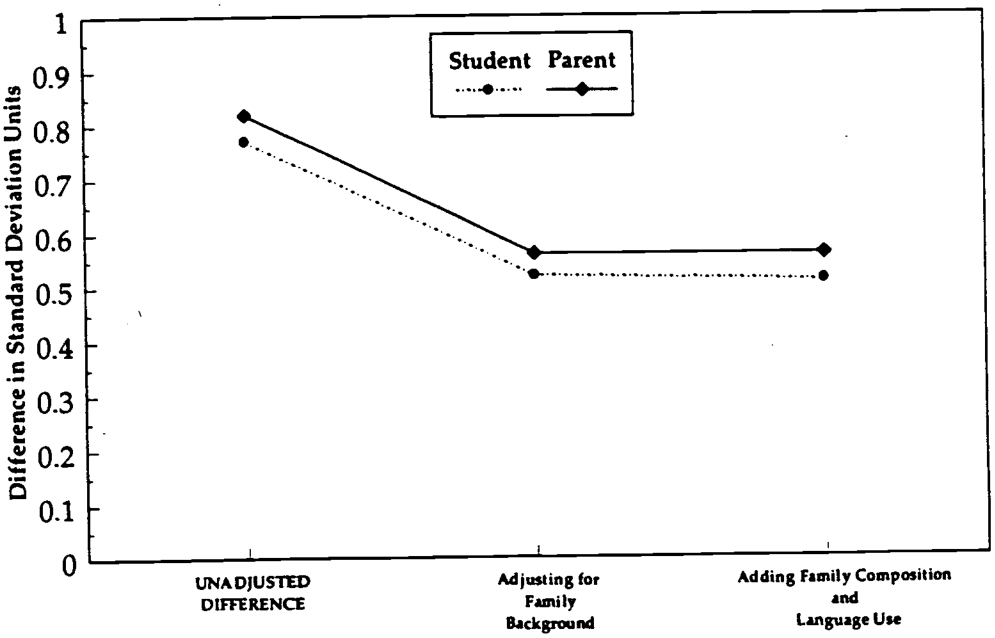

Figure 13 -- Mathematics Differences Between African Americans and Whites, Unadjusted and Adjusted Comparison of Grade 8 NELS Student and Parent Sources

The unadjusted differences between blacks and whites are quite similar across sources (.77 SD in the student data and .82 SD in the parent data; Figure 13); this difference reflects only disparities in students' and parents' identification of population-group membership. More importantly, the impact of adjusting for social context measures (family background, family composition, and language use to) is essentially identical regardless of the source of data (indicated by the parallel lines in the two regressions in Figure 11). In fact, the percent reduction differed by no more than two percentage points between the parent- and student-based data. Similar findings were obtained for twelfth grade and for the differences between Hispanics and whites (see Berends, Koretz, and Lewis, 1994).

For several reasons, however, these findings do *not* indicate that student reports are generally sufficient. Most important, the similarity of our regressions based on student and parent information is to some degree simply a result of the subset of variables we were able to include. That is, the subset of social context variables for which information was available from both parents and students included some for which agreement between students and parents was high, such as measures of family composition (two-parent household and number of siblings) and language use. Although parent and student responses were less consistent for measures of mother's and father's education, occupation, and income, these inconsistencies -- in the context of more consistent responses on other variables -- were not sufficient to cause substantial differences in the regression models.

Second, NAEP also measures student performance at grade four, and as one would expect, studies show that fourth graders' reports are suostantially less consistent with parental reports than are those of older students (Looker, 1989; Mason, Hauser, Kerckhoff, Foss, and Manton, 1976). Comparable analyses of grade four (which neither NELS nor HSB included) might therefore show greater discrepancies between parent-based and student-based analyses. Finally, NAEP data can be used to address a wide range of questions, and other analyses that require different background variables might be more affected by the reliance on student self-reports.

DISCUSSION

The adequacy of NAEP's measurement of social context depends on the purposes to which the data will be put. Data that are adequate for one purpose may be inadequate for another. Moreover, surveys such as NAEP are subject to numerous constraints, such as resource limitations and limits on the time participants are willing to contribute. Thus, devoting additional resources to improve measurement for one purpose is likely to weaken the survey for other purposes.

This study was premised on the assumption that one important purpose of NAEP is to provide data on the relative achievement of racial and ethnic minorities and that to do so without reference to the marked differences in the social context s of these groups would be misleading. Thus, the analyses reported here were tailored to assess the adequacy of NAEP's social-context measures *for the purpose of reporting population group differences in achievement*. For other purposes, NAEP's measurement of social context may be either less or more adequate than it is for this purpose (for specific examples of this, see Berends and Koretz, in press).

With this focus, we examined both the adequacy of NAEP's selection of social-context constructs and the adequacy with which they were measured -- primarily, the quality of the student reports upon which NAEP largely relies for information on social context.

The analyses reported here show that there are substantial weaknesses in NAEP's set of social-context measures and that they have an important practical effect: they lead to an appreciable overestimate of the differences in performance between students from different population groups but similar social contexts. A major cause of this problem is the omission from NAEP of a number of social context variables that are important influences on the disparities among population groups. As our comparisons to HSB and NELS show, additional information about parental occupation, household income, number of siblings, and ability grouping arrangements could improve adjusted estimates of test score differences among groups in the NAEP.

It is important to bear in mind the one exception to this generalization: Hispanic-white achievement differences in grade twelve, where controlling for social context led to *greater* decreases in NAEP than HSB. We can only speculate about the reason for this exception, but it seems plausible that it may stem from the dramatic growth and changed composition of the Hispanic population between 1980 (HSB) and 1990 (NAEP) (Hernandez, 1993; McArthur, 1993; McDonnell and Hill, 1993; National Center for Education Statistics, 1993).

In contrast, the results of our analyses of the quality of measurement -- of social context -- were equivocal. There is clear evidence from NELS and HSB, as well as from a variety of other sources, that students, even those as old as high school seniors, do not always provide information about social context consistent with that provided by their parents. The consistency between parents and students varies dramatically across variables. Some obvious family and individual characteristics -- such as the number of parents in the household, number of siblings, language use, and population group -- are reported quite consistently. At the other extreme are variables such as parental occupations and family income, about which students and parents typically report inconsistent information. No nationally representative surveys allow an evaluation of the adequacy of the social-context information provided by NAEP's fourth-grade cohorts, but the more limited available data suggest, as one would expect, that data from elementary school students are markedly less consistent with parental reports.

We were able to judge the practical importance of inconsistencies between students and their parents only for grades 8 and 12 and only for the subset of relevant social context variables for which NELS and HSB obtained information from both students and their parents. *For those specific variables and for the specific questions addressed here,* reliance on student reports has little practical effect. However, this subset of measures includes some for which student and parent responses were relatively consistent. Therefore, this finding need not imply that inconsistencies in the reporting of other social context variables would be similarly unimportant, that such inconsistencies would be unimportant for other analytical purposes, or that the typically larger inconsistencies at younger ages would have similarly small impact.

Three limitations of the analyses reported here suggest that our findings may understate NAEP's weaknesses for the specific purpose of reporting group differences independently of social context. First, HSB and NELS were used as benchmarks to compare the NAEP because they included richer sets of social context measures and were less dependent on student self-reports. But HSB and NELS also have their limitations for these purposes. For instance, neither provides adequate information on long-term poverty, neighborhood characteristics, and parenting practices in the home, all of which are significant in explaining the achievement differences between minority and non-minority students (Brooks-Gunn, Klebanov, and Duncan, in press; Brooks-Gunn, Duncan, Klebanov, and Sealand, 1993; for a review see Berends and Koretz, in press).

Second, because of the purposes of the study, we examined social-context variables that have been shown to predict achievement without regard to their likely measurement quality. If some of these variables are in fact poor indicators of social context they purport to measure, our findings would be distorted. For present purposes the most important example is probably our inclusion of NAEP's size-and-type-of-community variable (STOC), which made NAEP appear more adequate than it otherwise would. STOC is a powerful predictor of performance and substantially reduced the achievement differences between population groups in our analyses. However, this variable depends on unvalidated and questionable principal reports and does not correspond well with Census data (Lipmann, 1993). Indeed, after the present study was undertaken, NCES ended its use of STOC in reporting NAEP results because of concerns about its validity. (A replacement has not yet been devised.)

Third, as noted above, for want of an appropriate database to which to compare NAEP, we did not examine the fourth grade level, at which NAEP's reliance on student reports is likely to be more problematic (Looker, 1989). Until NAEP conducts a parent survey for its fourth grade sample or fourth grade is included in another nationally survey with both parent and student components, there will be no representative data with which to assess the adequacy of social context measures for younger children.

It is also important to note one limitation the impact of which is unclear. The operationalization of social context measures, including the phrasing of questions and their division into categories, was beyond the scope of our analyses but could have important effects on the adjusted test score differences and could have contributed to the differences observed among the databases. For example, while all three databases included measures of language use, NELS contained more detailed questions than HSB and especially NAEP. Additional studies of these aspects of measurement quality might further improve NAEP's social context measures.

The larger question is whether the purpose that motivated this study is one NAEP should serve: should NAEP report population-group differences jointly with information about social context? This could be done in numerous ways: reporting group differences after controlling for social context differences, as we did with our regression analyses; reporting unadjusted scores for groups broken down by social context variables as well as population group; or by reporting overall adjusted scores for population groups along with descriptive information about their social contexts. The adequacy of any of these approaches depends on having the appropriate social construct measures included and measured adequately.

This question, which has recently been debated in setting policy for NAEP reporting, affords no unambiguous answer. Some argue for adjustment, claiming that reporting unadjusted scores without reference to social context differences is inherently unfair, not very informative, and potentially very misleading. Others argue against adjustment, maintaining that adjusting for differences in social context (or reporting group differences along with corollary information about social context) sends an unacceptable message about educational standards. They contend that reporting without adjustment fc: social-context differences is necessary in order to communicate that similar expectations are held for all students, not only the privileged.

Even if one accepts that population-group differences should be reported in conjunction with information about social context, there is no fully sufficient way to do so. A simple approach that would appear to avoid the vexing question of how to adjust scores would be to present unadjusted group differences (or group means) along with corresponding information about social context differences. This approach, however, leaves the consumer of information with the impracticable task of disentangling the variables and interpreting what impact social context differences might have. A second approach would be to report scores for groups defined by both population group and social context variables, much as NAEP now reports state means overall and separately for population groups. However, this could only be done for one or two variables at a time; otherwise, results would become too numerous. Moreover, even breaking down scores only by a few variables at a time could lead to unreliable estimates based on few cases. Most important, because it would be confined to a few variables at a time, this approach is unlikely to solve the problem of misleading inferences posed by simple univariate reporting.

The third approach would be to report scores adjusted by a multivariate model, presumably in combination with unadjusted scores. Even if significant efforts were made to improve NAEP's data on social context, however, there would be several serious obstacles to the reporting of adjusted scores. First, the appropriate adjustment would depend on the question being asked. For some purposes, for example, one might want to control for school organization and educational practices, while for others one would not. Even for a given question, however, there is generally not a consensus in the educational research community on what the appropriate aajustments should be. Finally, resources are finite, and even a substantially improved measurement of social context would omit some constructs and leave others weakly measured.

Whether adjusted scores should routinely be presented in official NAEP reports despite these difficulties is a policy question that we do not presume to answer. However, we suggest that NAEP should be capable of yielding reasonably high-quality adjusted scores regardless, and the results reported here suggest that improvements in the database will be needed if it is to serve that purpose. Data can influence policy and practice through many channels beyond official reports of the statistical agencies that produce them (e.g., Coleman *et al.*, 1966; Coleman, Hoffer, and Kilgore, 1982). Secondary analysis of nationally representative data has often had a substantial impact on debates about policy and practice, and the Education Department has invested considerable resources in efforts to encourage secondary analysis of NAEP. Given the salience of issues of educational achievement in today's debates about education and NAEP's position as the pre-eminent and richest source of nationally representative information about student achievement, additional analysis of NAEP data could play a key role in shaping the debate. If NAEP is to play that role, it is hard to imagine an issue more important than helping to disentangle the relationships among population-group membership, social context, and achievement.

REFERENCES

Averch, H. A., Carroll, S.J., Donaldson, T.S., Kiesling, H.J., and Pincus, J. 1972. How effective is schooling? Santa Monica, CA: RAND.

Berends, M., and Koretz, D. In press. Identifying student at risk of low achievement in national data. Santa Monica, CA: RAND.

Berends, M., Koretz, D., and Lewis, E. 1994. Measuring racial and ethnic test score differences: Can the NAEP account for dissimilarities in social context? Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Borus, M. E., and Nestel, G. 1973. Response bias in reports of father's education and socioeconomic status. Journal of the American Statistical Association, 68: 816-820.

Brooks-Gunn, J., Duncan, G. J., Klebanov, P., and Sealand, N. 1993. Do neighborhoods influence child and adolescent development? American Journal of Sociology, 99: 353-395.

Brooks-Gunn, J., Klebanov, P., and Duncan, G. J. In press. Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics. Child Development.

Bryk, A. S., and Hermanson, K. L. 1993. Educational indicator systems: Observations on their structure, interpretation, and use. Review of Educational Research in Education, 19: 451-484.

Cohen, R. S., and Orum, A.M. 1972. Parent-child consensus on socioeconomic data obtained from sample surveys. Public Opinion Quarterly, 36: 95-98.

Coleman, J. S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., and York, R.L. 1966. Equality of educational opportunity. Washington, DC: U.S. Government Printing Office.

Coleman, J. S., Hoffer, T.B., and Kilgore, S. 1982. High school achievement. New York: Basic Books.

Coleman, J. S., and Hoffer, T.B. 1987. Public and private schools: The impact of communities. New York: Basic Books.

Duncan, O. D. 1961. A socioeconomic index for all occupations. In A. J. Reiss, Jr. (Ed.) Occupations and social status: 109-138. New York: Free Press.

Fetters, W. B., Stowe, P.S., and Owings, J.A. 1984. High school and beyond: A national longitudinal study for the 1980's: Quality of responses of high school students to questionnaire items. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Foertsch, M. A. 1992. Reading in and out of school. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Gamoran, A. 1987. The stratification of high school learning opportunities. Sociology of Education, 60: 135-155.

Gamoran, A., and Berends, M. 1987. The effects of stratification in secondary schools: synthesis of survey and ethnographic research. Review of Educational Research, 57: 415-435.

Gamoran, A., and Mare, R. D. 1989. Secondary school tracking and educational inequality: Reinforcement, compensation, or neutrality? American Journal of Sociology, 94: 1146-1183.

Hauser, R.M., Tsai, S., and Sewell, W.H. 1983. A model of stratification with response error in social psychological variables. Sociology of Education , 56: 20-46.

Hernandez, D. J. 1993. America's children. Washington, DC: Census Bureau.

Heyns, B., and Hilton, T.L. 1982. The cognitive tests for high school and beyond: an assessment. Sociology of Education, 55: 89-102.

Ingels, S. J., Abraham, S.Y., Karr, R., Spencer, B.D., and Frankel, M.R. 1990. National Education Longitudinal Studv of 1988: Data File User's Manual. Washington, DC: US Department of Education, National Center for Education Statistics.

Jencks, C. S., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., Michelson, S. 1972. Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books.

Jencks, C. S., Bartlett, S., Corcoran, M., Crouse, J., Eaglesfield, D., Jackson, G., McClelland, K., Mueser, P., Olneck, M., Schwartz, J., Ward, S., and Williams, J. 1979. Who gets ahead? The determinants of economic success in America. New York: Basic Books.

Jones, C., Clark, M., Mooney, G., McWilliams, H., Crawford, I., Stephenson, B., Tourangeau, R. 1980. High school and beyond 1980 senior cohort data file user's manual. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Johnson, E. G., and Allen, N.L. 1992. The NAEP 1990 technical report. Washington, DC: National Center for Education Statistics.

Kaufman, P. and Rasinski, K.A. 1991. National education longitudinal study of 1988: Quality of the responses of eighth-grade students in NELS:88. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kayser, B. D. and Summers, G.F. 1973. The adequacy of student reports of parental SES characteristics. Sociological Methods and Research, 1: 303-315.

Kerckhoff, A. C., Mason, W.M., and Sandomirsky Poss, S. 1973. On the accuracy of children's reports of family social status. Sociology of Education, 46: 219-247.

Koretz, D. 1991. State comparisons using NAEP: Large costs, disappointing benefits. Educational Researcher, 20(3): 19-21.

Koretz, D. 1992. What happened to test scores, and why? Educational Measurement: Issues and Practice, 11(4): 7-11.

Lee, V. E., and Bryk, A.S. 1988. Curriculum tracking as mediating the social distribution of high school achievement. Sociology of Education, 61: 78-94.

Lipmann, L. 1993. NAEP type of community variable. Unpublished memo, Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Looker, D. E. 1989. Accuracy of proxy reports of parental status characteristics. Sociology of Education, 62: 257-276.

Mare, R. D., and Winship, C. 1988. Ethnic and Racial Patterns of Educational Attainment and School Enrollment. In G. D. Sandefur and M. Tienda (Eds.), Divided opportunities: Minorities, poverty, and social policy: 173-203 . New York: Plenum Press.

Mason, W. M., Hauser, R.M., Kerckhoff, A.C., Sandomirsky Poss, S., and Manton, K. 1976. Models of response error in student reports of parental socioeconomic characteristics. In W. H. Sewell, R. M. Hauser, and D. L. Featherman (Eds.) Schooling and achievement in American society: 443-494. New York: Academic Press.

McArthur, E. K. 1993. Language characteristics and schooling in the United States, a changing picture: 1979 and 1989. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

McDonnell, L. M., and Hill, P. T. 1993. Newcomers in American schools: Meeting the educational needs of immigrant youth. Santa Monica, CA: RAND.

Mullis, Ina V.S., Dossey, J.A., Foertsch, M.A., Jones, L.R., and Gentile, C.A. 1991. Trends in academic progress. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Mullis, Ina V.S., Dossey, J.A., Owen, E. H., Phillips, G.W. 1991. The state of mathematics achievement. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Murnane, R. J. 1987. Improving education indicators and economic indicators: The same problems? Education Evaluation and Policy Analysis, 9: 101-116.

National Assessment of Educational Progress. (1988). Mathematics objectives 1990 assessment. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1989). Reading objectives 1990 assessment. Princeton, NJ: Educational Testing Service.

National Center for Education Statistics. 1993. Youth indicators 1993. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Oakes, J. 1990. Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science. Santa Monica, CA: RAND.

Oakes, J., Gamoran, A., and Page, R. 1992. Curriculum differentiation: opportunities, outcomes, and meanings. In P. W. Jackson (Ed.), Handbook of research on curriculum: 570-608. New York: Macmillan.

Rasinski, K. A., Ingels, S.J., Rock, D.A., and Pollack, J.M. 1993. America's high school sophomores: A ten year comparison. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Rock, D. A., and Pollack, J. 1991. Psychometric report of the NELS: 88 Base year test battery. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Sewell, W. H., Haller, A. O., and Portes, A. 1969. The educational and early occupational attainment process. American Sociological Review , 34: 82-92.

Sewell, W. H., Haller, A.O., and Ohlendorf, G.W. 1970. The educational and early occupational status attainment process: Replication and revision. American Sociological Review, 35: 1014-27.

Sewell, W. H., and Hauser, R.M. 1975. Education, occupation, and earnings: Achievement in the early career. New York: Academic Press.

Sewell, W. H., and Hauser, R.M. 1980. The Wisconsin longitudinal study of social and psychological factors in aspirations and achievements. In A. C. Kerckhoff (Ed.), Research in sociology of education and socialization: 59-100. Greenwich, CT: JAI Press.

Sewell, W. H., Hauser, R.M., and Wolf, W.C. 1980. Sex, schooling and occupational status. American Journal of Sociology , 86: 551-83.

Smith, M. S., and Scoll, B. W. 1995. The Clinton human capital agenda. Teachers College Record, 96: 389-404.

Table 1
Variable Definitions in NAEP, HSB, and NELS

VARIABLES	DEFINITION	NAEP	HSB	NELS
Test Scores				
Mathematics	Z-Score, weighted	X	X	X
Reading	Z-Score, weighted	X	X	X
Population Group	Mutually exclusive dummy variables for black, Hispanic, other, and white (reference)	X	X	X
Family Background				
Mother's Education	Dummy variables (effects coded) for less than high school, high school graduate, some college, college graduate, and missing data (reference)	X	X	X
Father's Education	Same as mother's education	X	X	X
Mother's Occupation	Dummy variables (effects coded) for 22 categories in NELS and 19 categories in HSB, including missing data category (reference)		X	X
Father' Occupation	Same as mother's occupation		X	X
Family Income	Dummy variables for 17 parent-reported categories in NELS and 7 student reported categories in HSB (low income categories as reference)		X	X
Family Composition				
Two-Parent Household	Dummy variable equal to one if student lived with both parents, 0 otherwise (reference)	X	X	X
Number of Siblings	Number of siblings (including step brothers and sisters), ranging from 0 to 5 or more		X	X
Language Use				
Other than English Spoken in Home	Dummy variables: usually speak English (reference), Spanish, and other language in the home; (Additional English and Spanish category available in NELS.)		X	X
Other than English Spoken Generally	Dummy variables: generally speak English (reference), Spanish, and other language; (Additional English and Spanish category available in NELS.)		X	X
Frequency of Speaking Language Other than English	Dummy variables for frequency of speaking a language other than English in the home -- never (reference), sometimes, and always	X		

Table 1 (continued)

VARIABLES	DEFINITION	NAEP	HSB	NELS
Community Characteristics				
Region of the Country	Dummy variables for Northeast, Central, West, and South (reference)	X	X	X
Locale	Dummy variables for urban, suburban, and rural (reference)	X	X	X
Size and Type of Community	Dummy variables for extreme rural, lower class metropolitan, upper class metropolitan (reference), urban fringe, main big city, medium city, and small place	X		
School Characteristics				
Percent on Free Lunch Program	Percentage of study body who receives free or reduced lunch	X	X	X
Private School	Dummy variable equal to one if private school, 0 otherwise (reference)	X	X	X
School Demographic Composition	Percentage of students who are black and percent who are Hispanic	X	X	X
School Socioeconomic Composition	School mean SES (e.g., family income, parents' educational attainments, father's occupational status, and household possessions)		X	X
School Attendance	Attendance rate		X	X
School Mean Achievement	School mean mathematics or reading	X	X	X
Curricular Differentiation				
Track Placement	Dummy variables for college (reference), general, and vocational track levels	X	X	
Ability Group Placement in Mathematics & Reading	Dummy variables for high, middle (reference), low, not grouped, don't know group from student reports			X
Percent in Remedial Reading	Percent of student body who are in remedial reading	X	X	X
Percent in Remedial Math	Percent of student body who are in remedial mathematics	X	X	X
Percent in Talented and Gifted	Percent of student body who are in talented and gifted program	X	X	X
Advanced English	Dummy variable equal to one if student is in advanced English, 0 otherwise (reference)			X
Advanced Mathematics	Dummy variable equal to one if student is in advanced mathematics, 0 otherwise (reference)			X

Table 2
Percent Agreement Between Students and Their Parents in Eighth and Twelfth Grades

	Grade 8 (NELS)	Grade 12 (HSB)
Population Group		
Race/Ethnicity	92	91
Family Background		
Mother's education	54	72
Father's education	49	70
Family income	NA	31
Mother's occupation	43	52
Father's occupation	44	54
Family Composition		
Two-parent household	90	92
Number of siblings	80	82
Language Use		
Language usually spoken	89	NA
Language usually spoken in the home	87	97

Notes: The information for this table was derived from Fetters, Stowe, and Owings (1984) and Kaufman and Rasinski (1991) as well of our own analyses of these items in NELS and HSB.