Using Test Content to Address Trend Discrepancies Between NAEP and California State Tests[1]

Xin Wei
Stanford University

Xuejun Shen
Stanford University

Brian Lukoff
Stanford University

Andrew Dean Ho
The University of Iowa

Edward H. Haertel
Stanford University

Paper presented at the meeting of the American Educational Research Association, San Francisco, California, April 2006

[1] This research was sponsored in part by a subcontract from the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), part of the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education. The first four authors of this paper contributed equally. Author names were listed in reverse alphabetical order.

Abstract

In 1998 and again in 2002, samples of eighth grade students in California were tested in reading as part of the state-level component of the National Assessment of Educational Progress (NAEP). In each of these years, all eighth graders in the state were also required to participate in the state's accountability testing, which included the reading test in the Stanford Achievement Tests, Ninth Edition (SAT-9). State-level comparisons of performance on these two assessments showed improvement in the SAT-9, but a slight decline on NAEP (not statistically significant). To examine whether this trend discrepancy might be attributable to content differences between the two tests, SAT-9 reading items were coded into categories corresponding to the NAEP content strands plus a category for items not aligned to the NAEP framework. Analyses of performance within strands indicate that content differences probably cannot explain the discrepant trends on the state accountability test versus NAEP, although differences related to item format are a strong possibility. Implications and alternative explanations are discussed.

Using Test Content to Address Trend Discrepancies
Between NAEP and California State Tests

1 Introduction

1.1 Confusions Resulting from Disordinal Trends

Statewide accountability testing results always make headlines. Positive test score trends are interpreted as showing an improvement in the quality of public education, an increase in student learning, and evidence of educational policies functioning as intended. Under the mandates of P.L. 107-110, the federal No Child Left Behind Act of 2001 (NCLB), state accountability test scores have taken on significant consequences for schools, increasing the burden on the validity of test score interpretations. The high stakes for state test scores can be contrasted with the relatively low stakes for results from the National Assessment of Educational Progress (NAEP), which is designed to evaluate the conditions and progress of education at the state and national levels. It would be reassuring if state and NAEP results showed similar trends for the same subjects, grades and years. In fact, as the NCLB legislation was being drafted, there was much discussion of a possible "confirming" role for NAEP vis à vis state tests required under the new legislation. Even though there was no explicit provision for comparing gains and gaps on state assessments versus NAEP, the law did mandate NAEP participation for sampled districts as a condition for receipt of Title I funds. The Department of Education's Fact Sheet on the new legislation stated, "Under H.R. 1 [the No Child Left Behind Act] a small sample of students in each state will participate in the fourth- and eighth-grade National Assessment of Educational Progress (NAEP) in reading and math every other year in order to help the US Department of Education verify the results of statewide assessments required under Title I to demonstrate student performance and progress" (U. S. Department of Education, n.d.).

However, between 1998 and 2002, results from NAEP and the Stanford Achievement Tests, Version 9 (SAT-9)—the state accountability exam given in California at that time—showed strikingly different results. While reading scores for California eighth graders on the SAT-9 *increased*, implying that the reading skills of eighth-grade students in California were improving over that time period, reading scores for California eighth graders on NAEP *decreased*, giving the opposite indication about reading performance. Thus, the data are dramatic because of the confusion resulting from common interpretations of test score trends: students are better at reading and worse at reading; teachers are better at teaching and worse at teaching; and measurement-driven educational reform is or is not working to achieve desired learning outcomes. This dissonance reveals the flaws in common interpretations of test score trends even when stark score trend discrepancies are not found.

1.2 Theories of Score Discrepancies

Comparisons between NAEP and state assessment results have appeared in the research literature (e.g., Linn, Graue, & Sanders, 1990; Koretz & Barron, 1998) and on occasion have received considerable media attention (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000; Bennett & Finn, 2000). As with California, it is typically found that state assessments offer a more positive picture than NAEP of gains over time. Three broad hypotheses may help explain such discrepancies, related to differences in the populations sampled by the two assessments, the conditions of their administrations (especially motivation), or the content tested (Koretz, McCaffrey, & Hamilton, 2001).

1.1 Differences in Sampling Frames

If examinee sampling frames for the state and NAEP assessments differed, or changed differentially over time, comparability of results would be compromised. All students enrolled in public schools in California must take tests annually under the STAR (Standardized Testing and Reporting) system, with certain narrowly defined exceptions. State NAEP results are based on a representative sample of roughly 200 public schools. In 1998, participation was voluntary, but sampling weights were adjusted to account for nonresponse. As noted above, participation was mandatory for schools in districts accepting Title I funds in 2002. Student participation within sampled schools is voluntary, but the proportion of students opting out of testing is very small. As with all assessments, there are policies permitting certain exclusions for some English Learners (EL students) and some students with Disabilities (SD) who cannot meaningfully participate even with a permissible accommodation. Thus, except for possible differences in exclusion policies, state and NAEP sampling frames should be highly similar

Permissible accommodations for students with disabilities or exclusion policies for English Learners may be different for NAEP versus state tests. Exclusions of EL or SD students change the definition of the population sampled. Thus, if different subsets of students are excluded from state versus NAEP assessments, this will bias achievement comparisons. In the present case, if exclusion policies were stable from 1998 to 2002 for each of the two respective assessments, then differential exclusion policies would be unlikely to affect *trend* comparisons. Trend comparisons would be distorted only to the extent that the groups of students included in one assessment but not the other showed different trends from those groups represented in both assessments. Because excluded groups are small relative to included groups, such effects would be unlikely to give rise to gross disparities. If, however, the exclusion rates for SD or LEP students were *not* constant over time, the effects of differential exclusion over time might distort measured trends dramatically (Haertel, 2003).

In 1998, four percent of California students were excluded from the NAEP sample (with accommodations) in connection with their designation as English Learners and/or students with disabilities. This percentage was unchanged in 2002.[2] Over this period, inclusion rules for California's accountability testing became *more* stringent in response to the requirements of the No Child Left Behind Act. That is, NCLB capped the percent of students who could be excluded for any reason. Because a lower exclusion rate would be expected to yield a lower mean score, any bias due to change in exclusion on the state assessment would probably be in the opposite direction from the trend discrepancy observed. Thus, different exclusion policies can probably be ruled out as an explanation for NAEP versus state assessment trend discrepancies.

1.2.2 Differential Student Motivation

Trend discrepancies may reflect differential changes over time in students' motivation. Many studies report that students lack motivation to perform well on low-stakes tests such as NAEP, because NAEP scores have no direct consequences for schools, teachers, or students (Karmos & Karmos, 1984). Therefore, NAEP scores may underestimate students' achievement levels. O'Neil (1992) compared students' performance under four motivational conditions:

[2] Data available at http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/S0092.asp?tab_id=tab3&subtab_id=Tab_1&printver=Y#chart, downloaded March 9, 2006.

financial reward, competition, personal accomplishment and standard NAEP test instructions. The results indicated that financial reward improved test scores for eighth graders but not for twelfth graders. In a later study, the same authors reported that external financial rewards could motivate eighth-grade students to put more effort into doing the test, which resulted in an increase in test scores on easier items but not on more difficult items (O'Neil, Sugrue, & Baker, 1996). However, a similar pattern was not found for twelfth-grade students. Because students' motivation plays an important role in test performance, different motivation levels of students on the NAEP test and state tests could help to explain test scores discrepancies at any one point in time. Note, however, that motivational differences between NAEP and state examinations would not give rise to trend discrepancies unless the motivation effect varied over time for one test or the other.[3]

The authors could not identify any plausible rationale for changes in eighth-grade students' motivation to perform well on the NAEP tests from 1998 to 2002. It is possible, however, that school personnel increased their efforts to encourage maximal student effort on the SAT-9 between 1998 and 2002, in response to the increased stakes associated with the test following the implementation of California's Academic Performance Index (API) in response to the Public Schools Accountability Act of 1999 and, later, the implementation of NCLB in 2002. Thus, differential change in motivation may have contributed to the observed score trend discrepancy.

1.2.3 Differences in Test Content or Format

Test content or format may differ between state and national tests. As Haertel (2003) observed, state tests are likely to be more closely aligned to state standards and curriculum than are the NAEP tests. In an item-level analysis comparing results for two tests in Chicago, Jacob (2002) reported distinct trends for different mathematics skills, but similar trends across reading content strands. He found that the improvement in math scores is largely driven by the questions of computation and number but very little by questions such as estimation, data interpretation and multiple-step problem-solving. Computation and number questions may account for a relatively larger proportion of high-stakes (state) test items than of the low-stakes (NAEP) test items. If teachers aligned their instruction to the high-stakes test items, students would be expected to show greater gains on computation and number questions versus estimation, data interpretation and multiple-step problem-solving questions. Thus, content differences together with incentives to align instruction with high-stakes state tests could account for the different trends observed in state versus NAEP test scores.

Muthén, Khoo and Goff (1997) showed that gaps on the 1992 NAEP Mathematics assessment are different across content strands. The subgroup differences on NAEP vary across content areas and test formats. Thus, specific test content may disadvantage certain subgroups but advantage others. For example, male students did better than female students on 1992 NAEP Grade 12 multiple-choice items, but worse on constructed-response items in the data analysis and

[3] If motivation on one test were so low that it affected test reliability, resulting in increased observed-score variance relative to true-score variance, then the magnitude of change over time expressed as an effect size would be attenuated. However, such attenuation alone could not give rise to *disordinal* trends on the two assessments. Moreover, low reliability should not affect NAEP estimates because NAEP's multiple imputation ("plausible values") and conditioning methodology yields estimates of score distributions in the metric of the underlying latent trait, not an observed-score metric.

statistics content strand. They concluded that test score discrepancies might be due to the different content mix, content weights and format of NAEP versus state tests.

1.3 Research Questions

The above three hypotheses are neither mutually exclusive nor exhaustive. As summarized in an editorial by William J. Bennett and Chester E. Finn, Jr. (2000), "It is normal for state tests to show better results than national ones. There are straightforward reasons for this: state tests are more narrowly designed; they test more basic skills; they intentionally align themselves to the state standards and curriculums (which national tests do not); and they provide more incentives, like grade promotion, for students to do well."

This paper offers a detailed case study of the hypothesis that content differences can account for discrepancies in NAEP versus state assessment score trends. Since NAEP and SAT-9 reading test contents are partially overlapping, the question arises as to whether performance trends on SAT-9 items within a particular content strand might match NAEP trends within that same strand, even if overall SAT-9 and NAEP trends differ.

A file of item-level responses of all eighth-grade students in California in 1998 and in 2002 was obtained from the State of California to enable such comparisons. Because California used the identical test from 1998 to 2002, and because item-level response data were available, it was possible in this study to track performance on specific aspects of reading performance without the need for linking or equating that would arise if test forms had changed. For NAEP, scaling is carried out separately within the three content strands, or "contexts," of Reading for Literary Experience, Reading for Information, and Reading to Perform a Task. SAT-9 composites aligned to each of these three strands were created. In addition, the SAT-9 also includes a "Reading Vocabulary" section that NAEP does not have, and some items in the SAT-9's "Reading Comprehension" section are more accurately coded as parts of a fifth "Other" strand than as one of the three NAEP content strands. To the extent that these different kinds of items measure different constructs, changes in student proficiency on SAT-9 Reading Vocabulary or Other items might not be reflected in NAEP results. Thus, these data offer a nearly unique opportunity to examine a question of importance to both educational measurement and standards-based reform policy.

2 Methods

2.1 Data Sources

We obtained 1998 and 2002 NAEP data from two sources. We made extensive use of the NAEP Data Explorer on the National Center for Education Statistics web site (http://nces.ed.gov/nationsreportcard/nde/), using it to gather all of the scaled-score data required for our analysis. We also obtained average raw scores for NAEP reading items in the three content strands for California eighth graders directly from ETS. These are nonpublic data, provided with the permission of the National Center for Education Statistics for use in this research.

Item-level SAT-9 data were provided by the State of California. This database contained right-wrong information for each item on the SAT-9 Reading Comprehension subtest, for every fourth- and eighth-grade student in California in 1998 and 2002 (see Table 2.1.1). In addition, the secure SAT-9 form T was obtained from the test publisher under a confidentiality agreement that prevents disclosure of specific content of items.

Table 2.1.1 Sample Sizes for California SAT-9 Data

	4th grade	**8th grade**
1998	432,353	399,618
2002	470,182	447,905

2.2 Coding SAT-9 Items

NAEP reading items are classified into three content strands: Reading for Literary Experience, Reading for Information, and Reading to Perform a Task. The National Assessment Governing Board's description of the reading framework (National Assessment Governing Board, 2002) notes that

> Many commonalities exist among the different reading contexts, including developing understanding, reflecting critically on the text, and analyzing the author's perspective. The contexts are not mutually exclusive….However, distinctions exist because various texts and tasks can place differing demands on the reader.…The contexts for reading and the reader's expectations may influence the comprehension process, determine what strategies and skills are used to develop meaning, and influence the extent to which content is integrated with prior knowledge.

Table 2.2.1 gives examples of passage types that items in each content strand might be associated with, and Table 2.2.2 gives a narrative description of each content strand.

Since the sets of passage types in Table 2.2.1 are non-overlapping, we initially coded each SAT-9 item based on the type of passage with which it was associated. For example, if one of the passages was a short story and it was followed by six questions, we coded all six as Reading for Literary Experience. After making this initial coding, we then reexamined each item in light of the content strand descriptions in Table 2.2.2. If an item clearly did *not* fit the description of the content strand, then we reclassified the item as Other.

It is clear that our results are highly dependent on an accurate classification of items into content strands. We wanted to avoid erroneously classifying items into one of the three content strands, so an item was classified in one of the three content strands if and only if two raters classified the item in the content strand of the parent passage (and the two raters agreed on the content strand of the parent passage). If *either* of the raters believed that an item did not fit the description of the appropriate content strand (or if the raters disagreed on the parent content strand), then the item was classified as Other.

For each set of items (fourth-grade and eighth-grade), a second team of two raters coded each passage and then each item according to the procedure above. This team's code for an item was a NAEP content strand (Reading for Literary Experience, Reading for Information, or Reading to Perform a Task) if both raters agreed; otherwise, the code was Other. On the fourth-grade items, the inter-team agreement on the overall passage codes was 100% and the agreement on the individual item codes was 85%. On the eighth grade items, the inter-team agreement on the overall passage codes was 89% and the agreement on the individual item codes was 83%.

The fourth and eighth grade reading comprehension subtests on the SAT-9 each contained 54 items; of those, we classified 4 on the eighth grade subtest into the Other category and we classified 14 on the fourth grade subtest into the Other category.

Table 2.2.1 Passage Types for each NAEP Reading Content Strand

Reading for Literary Experience	**Reading for Information**	**Reading to Perform a Task**
novels short stories poems plays legends biographies myths folktales	textbooks primary and secondary sources newspaper and magazine articles essays speeches	charts bus or train schedules, directions for games or repairs classroom or library procedures tax or insurance forms recipes voter registration materials maps referenda consumer warranties office memos

Source: National Assessment Governing Board, 2002

Table 2.2.2 NAEP Reading Content Strand Descriptions

Reading for Literary Experience	The reader brings his or her experiences and knowledge to the text in such activities as anticipating events, picturing settings, predicting consequences, analyzing actions, and considering the language of literary works. The reader thinks about the perspective of authors and characters and considers the language and story structure when reading for literary experience. Various types of texts are associated with reading for literary experience, including novels, short stories, poems, plays, legends, biographies, myths, and folktales.
Reading for Information	Reading for information is most commonly associated with textbooks, primary and secondary sources, newspaper and magazine articles, essays, and speeches. Some features that distinguish informational text from literary text are organization and the way information is presented. Informational text is organized by topic and supporting details, whereas literary text is organized by the structure of a story, poem, or drama. Informational texts may have boldface headings, graphics, illustrations, and captions that signal importance in the text. However, some commonalities exist between literary and informational text and the skills and strategies required for reading each. Both require people to critically analyze the text, reflect on it, and draw conclusions. When reading for information, readers need to know the specific text patterns, or forms of organization (e.g., cause and effect, sequential order, comparison/contrast, opinion and supporting arguments), to develop understanding. People frequently have different purposes for reading text of this nature—for example, to find specific pieces of information, answer a question, or get some general information when glancing through a magazine article. Reading informational text calls for orientations to the text that differ from those used in reading for literary experience because readers are specifically focused on acquiring information. When people read for information, they may select parts of the text they need, rather than reading from beginning to end.
Reading to Perform a Task	When people read to perform tasks, they use their expectations of the purpose and structure of practical text to guide how they select, understand, and apply information. Practical text may include charts, bus or train schedules, directions for games or repairs, classroom or library procedures, tax or insurance forms, recipes, voter registration materials, maps, referenda, consumer warranties, or office memos. The reader’s orientation involves looking for specific information to do something. Readers need to apply information, not simply understand it. In this type of reading, readers are not likely to savor the style or thought in the texts as they might in reading for literary experience.

Source: National Assessment Governing Board, 2002

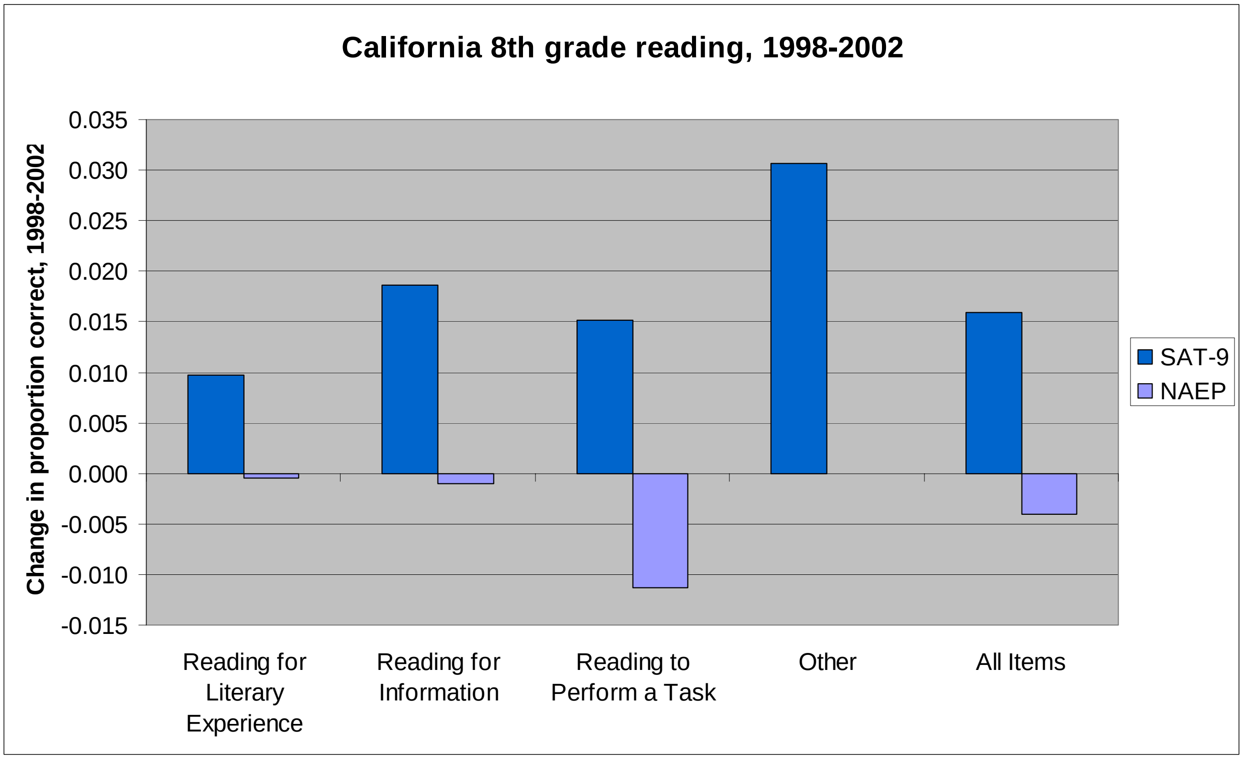
3 Results and Discussion

3.1 Comparison of average scores

The most direct comparison possible given the data available to us is between average *p*-values for SAT-9 item categories and average raw scores (expressed as proportion correct) for the corresponding content strands for NAEP eighth grade reading.[4] The rightmost set of bars in Figure 3.3.1 shows the disordinal trends between SAT-9 and NAEP—this is the same disordinal trend we observed earlier between SAT-9 average *p*-values and NAEP average scaled scores, but this time on the same metric (mean proportion correct), so that the magnitudes of the changes are comparable. It is clear from the first three sets of bars on the left that arranging SAT-9 items into the NAEP categories does not resolve the discrepant trends: even within the three content strands, SAT-9 and NAEP still show disordinal trends.

One interesting point is the fact that the largest increase between 1998 and 2002 on the SAT-9 was found on the items categorized as "Other" (those that did not fit into one of the three NAEP content strands). However, as only four items were classified as "Other," one should not read too much into this discrepancy.

Figure 3.1.1: Comparison of changes in proportion correct for NAEP and SAT-9 (8th grade)



[4] The 1998 eighth-grade NAEP reading assessment item pool included 39 multiple-choice, 56 short constructed-response, and 12 extended constructed-response items. Items in these respective formats were scored on scales of 0-1, 0-2, and 0-3, respectively. Of these 107 items, 102 were included in the tabulations of identical items from 1998 and 2002 obtained from ETS. Raw score means for each content strand and for the entire pool are expressed as proportions of the maximum possible score for that set of items.

The question naturally arises as to whether the observed discrepancies across NAEP strands and/or across SAT-9 categories might reflect no more than statistical noise. No precise statistical test is available to evaluate the statistical significance of NAEP differences, but two approximate tests suggest that the observed differences in the changes across NAEP content strands are probably not statistically significant. First, the significance tests available using the NAEP Data Explorer indicate that there is no statistically significant difference between 1998 and 2002 California eighth-grade NAEP reading scores for the overall reading scale or for any of the three strands. The NAEP Data Explorer does not permit a direct test of the interaction between strand and year. However, inasmuch as none of the strand-level changes is significantly different from zero and since all are in the same direction, it is highly unlikely that they differ significantly from one another.[5] Second, turning from the scale-score metric to the proportion-correct metric, ETS actually provided two distinct estimates of the 2002 mean proportion correct by strand, one from the operational NAEP data collection and the other from a smaller, concurrent equating study conducted in connection with the standardization of the NAEP exercise booklet design that year. From these two estimates, it is possible to calculate a single-degree-of-freedom estimate of the standard error of the mean proportion correct for each strand. This standard error, in turn, can be used to estimate the standard error of the 1998 proportion correct.[6] These two standard errors were then combined to obtain a standard error for the difference between 1998 and 2002 values. T ratios (with one degree of freedom) were formed by dividing each 1998-2002 raw proportion correct change by the corresponding standard error. The significance levels for these t-ratios for the three strands were .98, .53, and .18, all well above the nominal .05 level. For the entire set of items, the corresponding significance level was .52.

For the SAT-9 items, because we had access to complete data, the statistical significance of the observed differences across the category-level trends in Figure 3.1.1 could be examined directly. Formally, this requires a test of the year-by-category interaction in proportion-correct (mean p-values). We used a two-way ANOVA with repeated measures on one factor. The unit of analysis is students. Content strand is treated as the within-subjects effect and the year is treated as the between-subjects effect. The dependent variable is proportion-correct by content stand. The proportion-correct changes for content strands are treated as repeated measures for each student. Results of this analysis are based on a random sample of 4000 eighth grade students from the population. Because the sphericity assumption is violated, the Greenhouse-Geiser adjustment is used. Table 3.1.1 shows the results of analysis.

[5] Because random errors due to examinee sampling would be expected to be positively correlated across strands within year, it is a mathematical possibility that a contrast among trends for different strands would be statistically significant. The authors judge this to be highly unlikely because strand-by-strand changes over time are so small relative to their respective standard errors.

[6] ETS provided mean numbers of examinees responding to each item, for the 1998 estimates, the 2002 operational estimates, and the 2002 equating estimates. For 2002, our best estimate was a weighted average across the operational and equating samples, the weights being proportional to the sample sizes. To obtain the standard error for 1998, the standard error for 2002 was multiplied by the square root of the ratio of the total sample size for 2002 and the sample size for 1998.

Table 3.1.1 Results of Two-Way ANOVA with Repeated Measures (With Greenhouse-Geisser Adjustment) for Grade 8

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
Between subjects					
Year	1.737	1	1.737	10.399**	0.001
Error	668.153	3998	0.167		
Within subjects					
Content Strand	72.603	2.062	35.212	997.764**	< 0.001
Content Strand ×Year	0.448	2.062	0.217	6.155**	0.002
Error	290.919	8243.396	0.035		

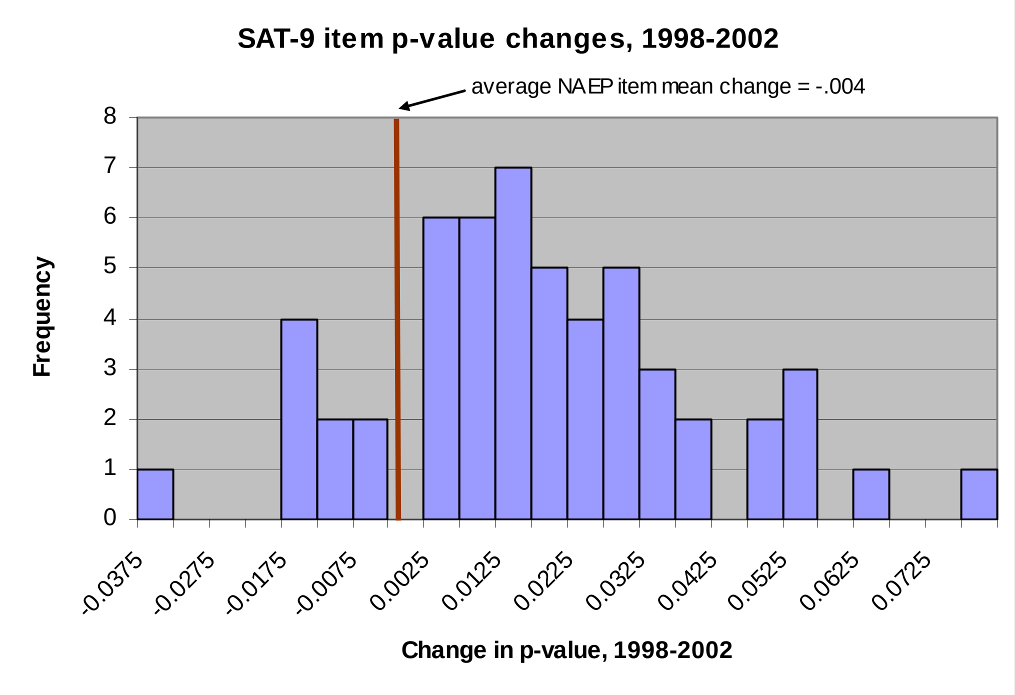
Measure: Proportion-Correct Scores
** p < .01.

The repeated-measures ANOVA reveals a significant between-subjects effect of year, F (1, 3998) = 10.40, p = 0.001, a significant within-subjects effect of content strands, F(2.06, 8243.40) = 997.76, p < 0.001, and a significant interaction effect between content strand and year, F(2.06, 8243.40) = 6.16, p = 0.002. The results indicate that there was significant improvement on the SAT-9 test scores from 1998 to 2002, students performed significantly differently in four content areas of reading test, and the change of proportion-correct scores from 1998 to 2002 varied significantly across content areas.

3.2 Would another content strand classification resolve the discrepancies?

The results from the previous section show that categorizing both SAT-9 and NAEP items into the content strands of the NAEP reading framework does not resolve the trend discrepancies between the eighth grade reading tests. A natural question to ask is whether a *different* classification of SAT-9 and NAEP items into content strands could resolve the trend discrepancy. Unfortunately, we did not have available to us individual NAEP item means, so we were unable to answer this question directly. However, we did know the individual SAT-9 item means (*p*-values, because all SAT-9 items were dichotomously scored) as well as the average change in NAEP item means. Figure 3.3.2 overlays a vertical line at the point of the average change in NAEP item means over a histogram of the 54 SAT-9 item *p*-values.

Figure 3.2.1: SAT-9 item p-value changes compared to average change in NAEP item mean



The location of the average NAEP mean item change line—splitting the SAT-9 items into groups of 9 items and 45 items—is visual evidence that it is unlikely that there exists another content strand classification scheme that would resolve the trend discrepancies within content strands.

To try to make this conclusion more precise, we ran a simulation study. Suppose there are G content strands, so that NAEP items are partitioned into groups N_1, N_2, …, N_G and SAT-9 items are partitioned into groups S_0, S_1, S_2, …, S_G, where S_0 is a group of $O = |S_0|$ items classified as “Other” (“non-NAEP-like”). Let y_1, y_2, …, y_n represent item-level trends for NAEP items (n is the number of NAEP items) and let x_1, x_2, …, x_{54} represent the individual p-value trends for SAT-9 items. The assumption that trends are the same across the two tests within content strands can be written as

$$\frac{1}{|N_i|}\sum_{j\in N_i} y_j = \frac{1}{|S_i|}\sum_{j\in S_i} x_j\,, \qquad \text{for } i = 1, 2, \ldots, G.$$

Under this assumption, we can deduce that

$$\text{NAEP trend} = \frac{1}{n}\sum_{g=1}^{G}\sum_{i\in N_g} y_i = \frac{1}{n}\sum_{g=1}^{G}\frac{|N_g|}{|S_g|}\sum_{i\in S_g} x_i\,,$$

where to get the second equality we used the assumption of equal group trends. In our simulation study we randomly partitioned the SAT-9 items into G groups (ignoring a random selection of O items classified as “Other”), computed the expression on the right side of the equation above, and then compared this value to the actual NAEP trend of -.004. We did this repeatedly and then calculated the proportion of the time that the value on the right side of the equation above was equal to or lower than the actual NAEP trend. If this proportion is low, then it is evidence that there is unlikely to be a meaningful selection of G content strands so that the

NAEP and SAT-9 trends align. We repeated the process for $G = 2, 3, 4$, and 5 and $O = 0, 5, 10, 15, 20$, and 25. In all cases, the proportion of the time that the expression on the right side of the equation above was at or below the NAEP trend was 1% or below.

3.3 Comparing SAT-9 p-values with NAEP scaled scores

Since SAT-9 average *p*-values and NAEP scaled scores are not on the same metric, it is imprudent to compare the magnitude of the 1998-2002 change in NAEP scaled scores with the change in SAT-9 average *p*-values. One notable exception to this rule is that we should not expect to see scores increase in one metric and decrease in the other if they are measuring the same thing, which is what we saw in the California eighth grade reading data.

Of course, we can compare changes in SAT-9 scores across subgroups and across content strands, and we can do the same for NAEP scores. By placing graphs next to each other, one for SAT-9 changes and one for NAEP changes—and ignoring the scale on each—one can get a sense of whether trends within subgroups and content strands are similar between the two tests.

Figures 3.3.1 and 3.3.2 show two separate sets of bar graphs of the same eighth grade reading data: in the first, the colors correspond to subgroups[7] and the groups of bars correspond to content strands; in the second, the colors correspond to content strands and the groups of bars correspond to subgroups. From these graphs it is clear that while on the whole the trends by content strand do not match up between NAEP and SAT-9 (as was clear from the direct comparison of average p-values in Figure 3.1.1), the trends are closer for some subgroups than for others. Black students—and to a lesser extent, Hispanic students as well—have SAT-9 score trends in the same direction as the corresponding NAEP trend for two out of the three content strands.

Figures 3.3.3 and 3.3.4 show the same two sets of bar graphs, this time for fourth-graders. Unlike the eighth grade data, the SAT-9 and NAEP score trends are both in the same direction (increasing from 1998 to 2002). However, while the SAT-9 scores generally increased about the same for both content strands (there were no Reading to Perform a Task items on the fourth grade NAEP), on NAEP scores for the Reading for Literary Experience content strand increased more than for the Reading for Information content strand. We also again see a difference in trends between subgroups: for White, Asian-American, and to a lesser extent Hispanic students, a larger increase on Reading for Information items on the SAT-9 is countered with a larger increase for Reading for Literary Experience on the NAEP. In contrast, Black students improved more on Reading for Literary Experience items on both tests. In summary, visual inspection of within-assessment patterns of changes by strand (category) and subgroup offer little evidence of meaningful content-related differences that might help to explain observed trend discrepancies.

[7] Both the SAT-9 and NAEP have subgroup categories for White, Black, and Hispanic. The California SAT-9 data has both Asian-American and Filipino groups, while NAEP has only an Asian-American group. Since NAEP's Asian-American group includes Filipino students, we combined the SAT-9 groups of Asian-American and Filipino for comparison purposes.

Figure 3.3.1: Trends for 8th grade reading, by content strand and subgroup

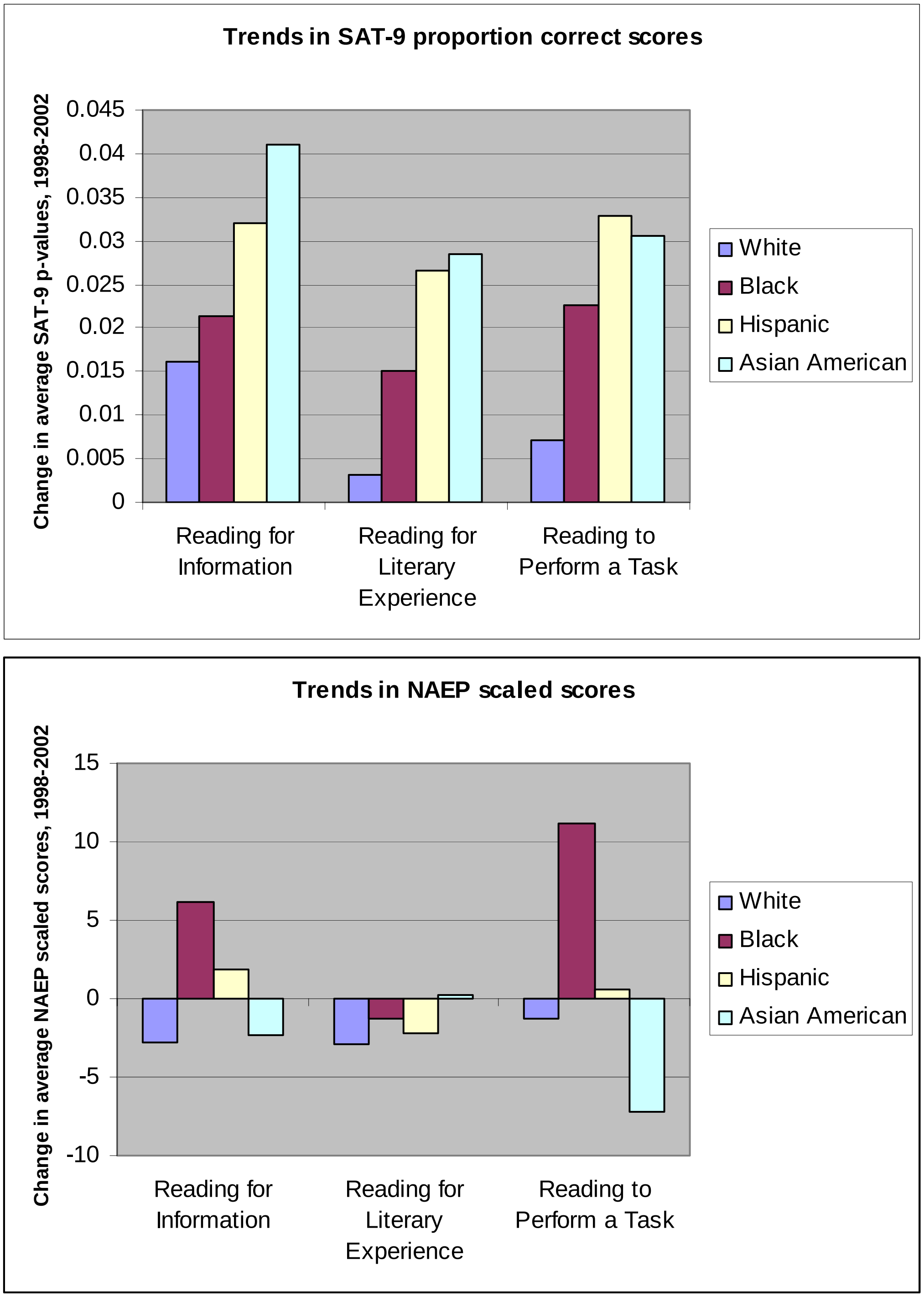
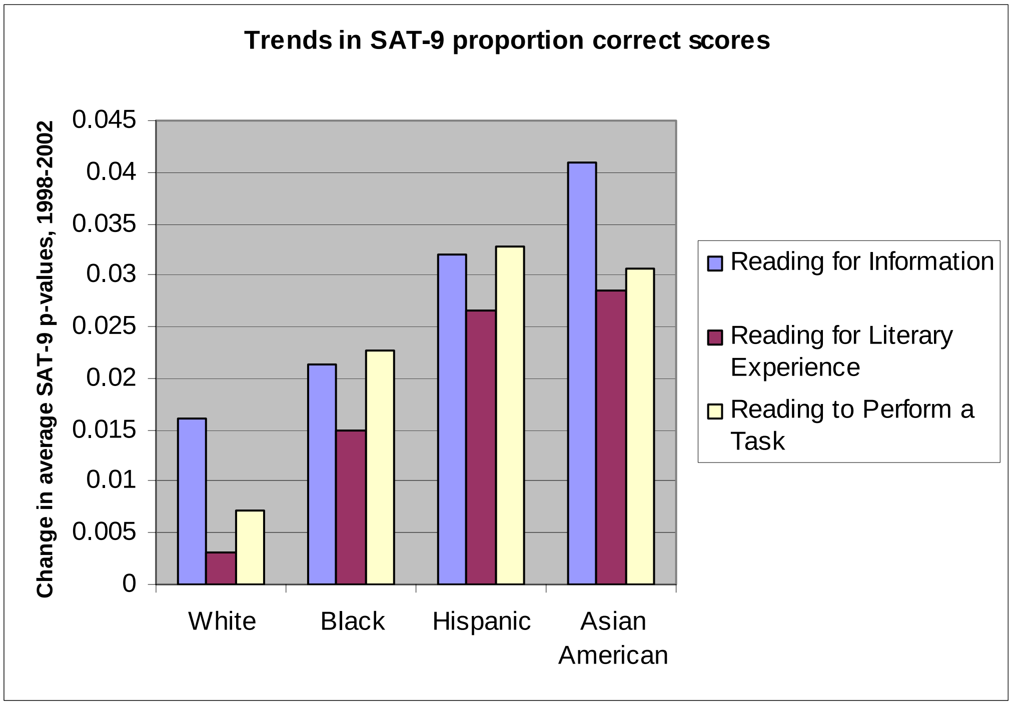


Figure 3.3.2: Trends for 8th grade reading, by subgroup and content strand



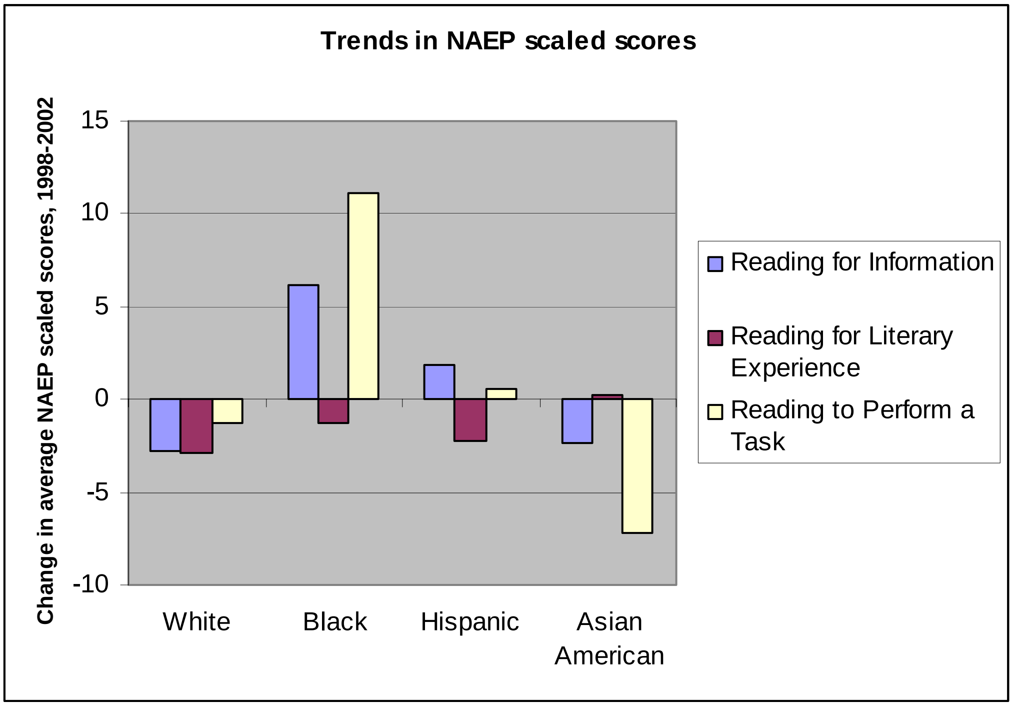


Figure 3.3.3: Trends for 4th grade reading, by content strand and subgroup

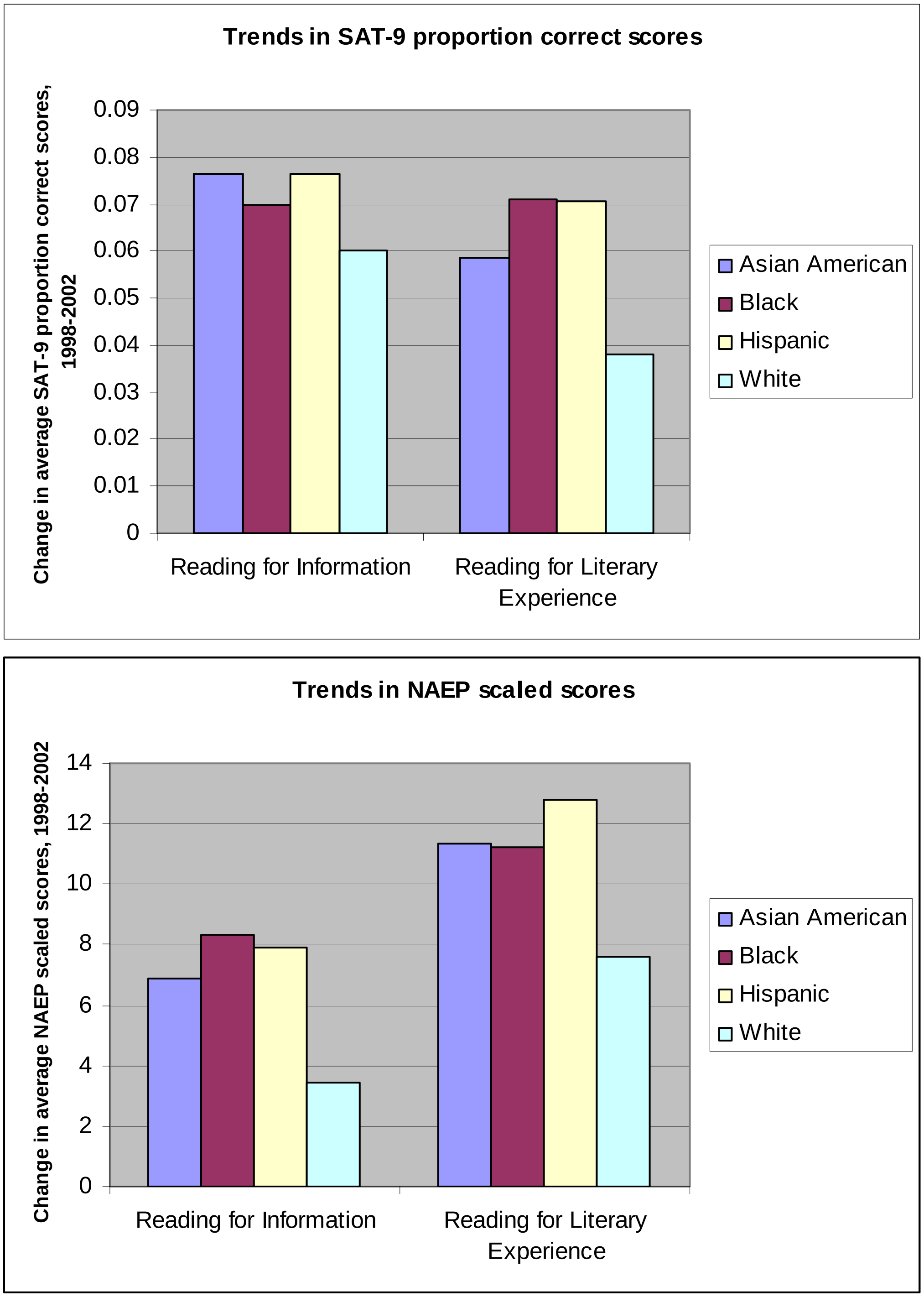
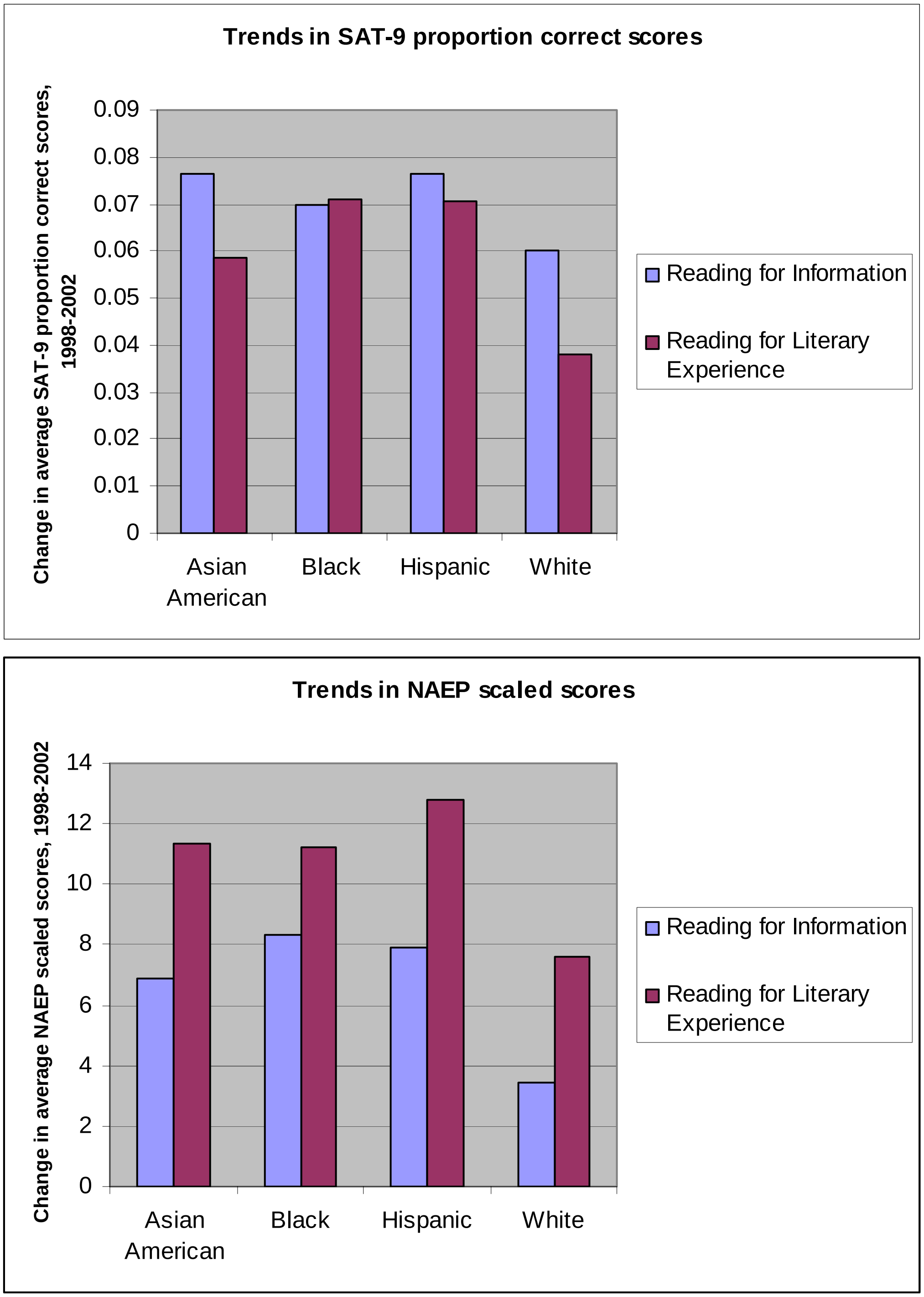


Figure 3.3.4: Trends for 4th grade reading, by subgroup and content strand



3.4 Two-way ANOVA for SAT-9 strands and demographic subgroups

Figures 3.3.1 through 3.3.4 appear to show that the trends of average SAT-9 item *p*-values vary for different subgroups but not for different content strands. To confirm this impression, we used a two-way ANOVA with repeated measures to test the effects of content strand, student ethnicity, and their interaction on the changes of SAT-9 item *p*-values from 1998 to 2002 for the fourth and eighth grades. Note that the ANOVA reported in Section 3.1 used total scores by strand as the dependent variable, and students as the unit of analysis. For the analysis reported here, items are regarded as random samples from universes of possible items of each category type. The student sample size (over 100,000) is treated as effectively infinite. Thus, sampling error due to students is ignored. The unit of analysis is the test item. Content strand is treated as a between-items effect, as the items are grouped by strands. The item *p*-value changes for the respective racial/ethnic groups are treated as repeated measures for each test item. The arcsine transformation is applied to the *p*-values to stabilize their variances (Kuehl, 1994), and the item *p*-value changes are thus calculated by the formula:

$$2\sin^{-1}\sqrt{2002\ \text{p-value}} - 2\sin^{-1}\sqrt{1998\ \text{p-value}}\,.$$

Mauchly's Test of Sphericity indicates that the sphericity assumption does not hold for these data for either grade level. Therefore the results reported below are adjusted with the Greenhouse-Geisser correction. Table 3.4.1 and Table 3.4.2 illustrate the ANOVA output for both grades.

Table 3.4.1 Results of Two-Way ANOVA with Repeated Measures (With Greenhouse-Geisser Adjustment) for Grade 4

Source	**Sum of Squares**	***df***	**Mean Square**	***F***	***p***
Between Items					
Intercept	4.209	1	4.209	266.622***	< 0.001
Content Strand	0.079	2	0.040	2.509	0.091
Error	0.805	51	0.016		
Within Items					
Ethnicity	0.067	1.533	0.044	25.088**	< 0.001
Content Strand x Ethnicity	0.015	3.065	0.005	2.726*	0.049
Error	0.136	78.168			

Measure: Transformed Item *p*-Value Change
* *p* < .05, ** *p* < .01.

Table 3.4.2 Results of Two-Way ANOVA with Repeated Measures (With Greenhouse-Geisser Adjustment) for Grade 8

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
Between Items					
Intercept	0.544	1	0.544	57.855**	< 0.001
Content Strand	0.032	3	0.011	1.143	0.341
Error	0.470	50	0.009		
Within Items					
Ethnicity	0.076	2.069	0.037	74.403**	< 0.001
Content Strand x Ethnicity	0.005	6.206	0.001	1.764	0.111
Error	0.051	103.441	0.000		

Measure: Transformed Item *p*-Value Change
** $p < .01$.

As shown in Table 3.4.1, for fourth grade reading, the within-items effect—student ethnicity—is highly significant, *F*(1.53, 78.17) = 25.09, $p < 0.001$. The between-items effect, content strand, is not significant, *F*(2, 51) = 2.51, $p = 0.09$; however, the interaction between ethnicity and content strand is just barely significant, *F*(3.07, 78.17) = 2.73, $p = 0.049$. The intercept of grand mean *p*-value change is significant with *F*(1, 51) = 266.62, $p < 0.001$. A similar pattern is found for the eighth grade reading items. Student ethnicity has a significant effect, *F*(2.07, 103.44) = 74.40, $p < 0.001$. The effect of content strand is not significant, *F*(3, 50) = 1.14, $p = 0.341$. However, the interaction between ethnicity and content strands is this time not significant, *F*(6.21, 103.44) = 1.76, $p = 0.111$. The intercept of grand mean *p*-value change is significant, *F*(1, 50) = 57.86, $p < 0.001$.

The ANOVA results indicate that there was significant improvement on the SAT-9 reading items from 1998 to 2002 for the fourth and eighth grades, and the mean *p*-value changes varied significantly across ethnic subgroups. However, there was no significant effect of content strand on the changes of item *p*-values. Only in fourth grade was the interaction between content strand and student ethnicity statistically significant, and even then the *p*-value just barely crosses the $\alpha = .05$ threshold. These results confirm the conclusion from the previous analyses, that examining item performance changes by content strand is unlikely to resolve the discrepant findings concerning California eighth-grade reading achievement trends on the SAT-9 versus NAEP.

4 Conclusions

Policy makers and the public may subscribe to the commonsense notion that "reading is reading is reading." That is, all objective tests of reading comprehension, based on reading brief passages and answering questions about is read, are testing essentially the same thing. This perception is reinforced by the typically high correlations observed among alternative tests within a subject area (e.g., reading or mathematics) and even across subject areas. As Koretz, et al. (2001), explain, however, similarities in rankings need not imply similarities in means (cf. also Klein, et al., 2000). Even tests that are highly correlated at any one point in time can show completely different patterns of changes in means from one year to another.

In this paper, we carefully examined data from two large-scale assessments showing such discrepant trends. Even though the California eighth-grade reading score decline on NAEP was not statistically significant, it differed markedly from the dramatic improvement shown for the same subject-grade combination on the SAT-9 (an effect size of .08). Three broad explanations for this trend discrepancy were considered. Differences in sampling frames appeared unlikely to account for the difference. If anything, more stringent SAT-9 inclusion policies under NCLB would have been expected to contribute to a trend discrepancy in the opposite direction. Motivational changes from 1998 to 2002 on the SAT-9 might help to explain the observed trend difference, but anecdotal reports at the time indicated that across the state, teachers and administrators encouraged students to do their best on the SAT-9 even in 1998, the year the test was introduced statewide in California. The major hypothesis considered involved possible content discrepancies between NAEP and the SAT-9. Available data on NAEP permitted examination of content effects down to the level of content strands. Graphical and statistical analyses for the eighth-grade public-school population as a whole and for major racial/ethnic subgroups, gave little or no evidence of content-strand-level effects that might account for the overall discrepancy observed. Thus, the source of the discrepancy remains unexplained.

Two important limitations of the current study must be noted. First, an ideal analysis would have used concordant subsets of items from each of the two assessments, setting aside not only SAT-9 items in an "Other" category but also NAEP items that showed a poor content match with those SAT-9 items that remained. Because item-level data from NAEP were not available, we were only able to compare SAT-9 item subsets to scale-score or raw-score performance on the three predefined subsets of NAEP items classified into the Literary-Experience, Information, and Perform-a-Task content strands. Second, and perhaps more significantly, although we were able to disaggregate NAEP items according to *content*, we were unable to disaggregate NAEP items according to *format*. As noted in passing, the NAEP eighth-grade reading exercise pool included roughly 35 percent dichotomously scored multiple-choice items, and just over 50% brief constructed response items scored on a 0-2 scale, plus about 12% extended constructed response items scored on a 0-3 scale. It is possible that eighth-grade performance on NAEP multiple-choice items tracked the trend observed for the SAT-9 multiple-choice items, but that performance on less familiar item formats declined sufficiently to give rise to the slightly negative overall trend.

The attempt to explain the score trend discrepancy is worthwhile, but the major implication for test interpretation and use remains the same regardless of whether this discrepant score pattern is attributable to content or format differences between NAEP and the SAT-9, to a motivation change from 1998 to 2002 specific to the SAT-9, to more extensive or more sophisticated training in test-taking skills, or to other factors (Koretz, 2005). The eighth-grade

SAT-9 data offer yet another illustration of the pervasive finding that score gains on high-stakes tests may fail to generalize to performance on other examinations. That is not to say that gains are entirely illusory, merely that they may often be overstated. Great caution is urged in the interpretation of score gains on high-stakes tests.

References

Bennett, W. J., & Finn, C. E., Jr. (2000, October 27). The Real Improvement in Texas Schools. *The New York Times*, p. A31.

Haertel, E. H. (2003). *Including Students with Disabilities and English Language Learners in NAEP: Effects of Differential Inclusion Rates on Accuracy and Interpretability of Findings* (Paper prepared for the National Assessment Governing Board). Retrieved July 28, 2005 from http://www.nagb.org/pubs/conferences/haertle.doc.

Jacob, B. (2002). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. Working Paper #8968, National Bureau of Economic Research (NBER), Cambridge, MA.

Karmos, A. H., and Karmos, J. S. (1984). Attitudes towards standardized achievement test performance. *Measurement and Evaluation in Counseling and Development, 12*, 56-66.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What Do Test Scores in Texas Tell Us?* (RAND Issue Paper No. IP-202). (Available at http://epaa.asu.edu/epaa/v8n49)

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (104th Yearbook of the National Society of Education, Part 2, pp. 99-118). Malden, MA: Blackwell.

Koretz, D. and Barron, S. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). Report, The RAND Corporation, Santa Monica, CA.

Koretz, D., McCaffrey, D., and Hamilton, L. (2001). Toward a framework for validating gains under high-stakes conditions. CSE Technical Report #551, University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

Kuehl, R.O. (1994). *Statistical principles of research design and analysis.* Belmont, CA: Duxbury Press.

Linn, R., Graue, M., and Sanders, N. (1990). Comparing state and district results to national norms: The validity of claims that “everyone is above average”. *Educational Measurement: Issues and Practice*, 9(3):5–14.

Muthén, B.O., Khoo, S.T., and Goff, G.N. (1997). Multidimensional description of subgroup differences in mathematics achievement data from the 1992 National Assessment of Educational Progress. CSE Technical Report #432, University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

National Assessment Governing Board. (2002). *Reading Framework for the 2003 National Assessment of Educational Progress*. Retrieved January 26, 2006, from http://www.nagb.org/pubs/reading_framework/toc.html

O'Neil, H. F., Jr. (1992) Experimental Studies on Motivation and NAEP Test Performance. Final Report. NAEP TRP Task 3a: Experimental Motivation. National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1996). Effects of Motivational Interventions on the National Assessment of Educational Progress Mathematics Performance. *Educational Assessment*, *3*, 135-157.

U. S. Department of Education. (n.d.). *Overview: Fact Sheet on the Major Provisions of the Conference Report to H.R. 1, the No Child Left Behind Act*. Retrieved July 28, 2005 from https://www.ed.gov/nclb/overview/intro/factsheet.html.