C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Raising the Stakes of Test Administration: The Impact on Student Performance on NAEP

CSE Technical Report 360

Vonda L. Kiplinger
CRESST/University of Colorado at Boulder

Robert L. Linn
CRESST/University of Colorado at Boulder

March 1993

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported in part under the National Center for Education Statistics Contract No. RS90159001 as administered by the U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the National Center for Education Statistics or the U.S. Department of Education.

RAISING THE STAKES OF TEST ADMINISTRATION:

THE IMPACT ON STUDENT PERFORMANCE ON NAEP

Abstract

It has been argued that NAEP provides an underestimate of student achievement because the assessment has no consequences for the students, their teachers, or for their schools. In contrast, it is hypothesized that since student performance on high-stakes tests has serious consequences for students or for teachers and schools, students have a higher "stake" in performing well. The purpose of this study is to investigate whether differences in test administration conditions and presumed levels of motivation engendered by the different testing environments affect student performance on NAEP administrations. The testing conditions under study are the "low-stakes" environment of the current NAEP administration and a higher-stakes environment typified by many state assessment programs. Two subsets of NAEP items were administered as part of Georgia's regular Curriculum-Based Assessments (CBA).

The results from Georgia's participation in the 1990 NAEP Trial State Assessment (TSA) provides the benchmark against which the state-embedded CBA results are compared. The means of the first subset of NAEP items are significantly higher in the 1992 CBA administration than in the 1990 TSA administration (effect size = 0.18). The TSA and CBA means for the second subset, however, are not significantly different. Similar results are obtained when male, female, white, and black student subpopulations are analyzed separately.

It may be that the difference in the results for the two subsets of items is a function of (a) their relatively difficulty; (b) contextual differences in the administration of the Block 7 NAEP items as part of the 1992 CBA as compared to the earlier administration in the 1990 TSA; (c) the additional 2-3 months of instruction received by students in the May 1992 CBA administration; or (d) real year-to-year differences in student achievement.

RAISING THE STAKES OF TEST ADMINISTRATION: THE IMPACT ON STUDENT PERFORMANCE ON NAEP

Summary

Recently, many questions have been raised concerning the effect of students' motivation on their performance on the National Assessment of Educational Progress (NAEP). It has been argued that NAEP provides an underestimate of student achievement because the assessment has no consequences for the students, their teachers, or for their schools; that is, to the individual student, the NAEP is a low-stakes test. In contrast, it is hypothesized that since student performance on high-stakes tests has serious consequences for students or for teachers and schools, students have a higher "stake" in performing well. Thus, the stakes are presumed to motivate performance commensurate with students' actual capabilities.

The purpose of this study is to investigate whether differences in test administration conditions and presumed levels of motivation engendered by the different testing environments affect student performance on NAEP administrations. The testing conditions under study are the "low-stakes" environment of the current NAEP administration and a higher-stakes environment typified by many state assessment programs. Items in the released Block 7 of the NAEP eighth-grade mathematics assessment were embedded in Georgia's state assessment, the Georgia Curriculum-Based Assessments (CBA), and useable data were obtained for 80,836 students. In order to reduce the burden on individual examinees, the block was split into two segments of the first nine items (administered to 40,403 students) and the last eight (administered to 40,433 students).

The results from Georgia's participation in the 1990 NAEP Trial State Assessment (TSA) provide the benchmark against which the state-embedded CBA results are compared. The analysis indicates that the state average percent correct values for the 1992 CBA and the 1990 TSA are highly correlated: $r = 0.97$. The means of the first nine NAEP items are significantly higher in the 1992 CBA administration than in the 1990 TSA administration.

The TSA and CBA means for the last eight items, however, are not significantly different. The significant difference on the first nine items represents an effect size of 0.18; the effect size for the last eight items, though not significant, is in the opposite direction (-0.04). Similar results are obtained when male, female, white, and black student subpopulations are analyzed separately. Item-level statistics also demonstrate a great deal of consistency in changes in percent correct from 1990 to 1992 for the total sample and for each of the four subpopulations. The splitting of the Block 7 items into two subsets of items and administration on different forms in the CBA provides a check on possible effects of contextual factors associated with administration of different forms. Small differences in the means of the NAEP item subsets as a function of test form suggest that there are small context effects stemming from the different orders of subject matter in the different test forms.

It may be that the difference in the results for the first nine and last eight items is a function of (a) their relatively difficulty; (b) contextual differences in the administration of the Block 7 NAEP items as part of the 1992 CBA as compared to the earlier administration in the 1990 TSA; or (c) real year-to-year differences in student achievement.

RAISING THE STAKES OF TEST ADMINISTRATION: THE IMPACT ON STUDENT PERFORMANCE ON NAEP

Vonda L. Kiplinger and Robert L. Linn*

University of Colorado at Boulder/CRESST

Introduction

Recently, many questions have been raised concerning the effect of students' motivation on their performance on the National Assessment of Educational Progress (NAEP). A number of people have argued that NAEP provides an underestimate of student achievement because the assessment has no consequences for the students, their teachers, or for their schools; that is, to the individual student, the NAEP is a low-stakes test. In contrast, it is hypothesized that a high-stakes testing environment, in which a student's performance has serious consequences for students (e.g., for grade promotion or graduation) or for teachers and schools (e.g., teacher recognition or allocation of funds), leads to greater achievement motivation and subsequent effort to score as high as possible on the test. Examples of high-stakes tests for individual students are the ACT and SAT, on which student score may be a determining factor in admission to U.S. colleges and universities, and the New York State Regents examinations, which can affect a student's high school grade point average and acceptance into college. Since such tests affect future educational opportunities and life chances for students, those who take the tests have a higher "stake" in performing well; thus, the stakes are presumed to motivate performance commensurate with students' actual capabilities.

Shanker (1990) compares performance on the high-stakes New York Regents exams with that on the low-stakes NAEP and notes that seemingly poorer performance on the NAEP tests may be due to the fact that "kids know the tests don't count." He acknowledges that high-stakes testing can have

* Ms. Kiplinger is a doctoral student in the School of Education, University of Colorado at Boulder. Professor Linn is Co-Director of the National Center for Research on Evaluation, Standards, and Student Testing at the School of Education, University of Colorado at Boulder.

negative effects such as increasing the likelihood of student cheating and teachers teaching what is likely to be on the test while neglecting other areas of instruction.[1] However, Shanker (1990) further argues that:

> . . . [it is] possible that low-stakes testing also has some serious disadvantages. If students know that what they do on a test doesn't matter, they may decide it's not worth their while to put forth any effort. And it could be that this explains the low level of achievement we have seen in NAEP examinations.

The stakes associated with a test can be increased when results have consequences for teachers or school administrators. The simple listing of results for individual schools in the newspaper, for example, can place considerable pressure on principals and teachers to assure that their students achieve high scores. Teachers, in turn, can increase the stakes for students by stressing the importance of their doing their best.

---

[1] A number of recent studies have questioned the validity of apparent gains in student scores on standardized tests. Allington and McGill-Franzen (1992), for example, demonstrate that increased use of high-stakes reading assessment in the primary grades in New York was accompanied by increases in average student performance *and* even greater increases in proportions of students retained or identified as handicapped *in the grade prior to the grade at which the high-stakes assessment occurred*. Thus, increases in retention and placement in special education may result in an artifactual effect suggesting that reading achievement is "rising in schools when, in fact, shifts in the patterns of cohort stability account for the higher reported reading levels" (p. 4).

Linn, Graue, and Sanders (1990), in a study that investigated the reported "Lake Wobegon Effect" that "everyone is above average" (i.e., all 50 states) on nationally-normed achievement tests, concluded that while test scores appear to have risen, the results may not be artificially inflated, as suggested by Cannell (1987). The apparent increase may be attributable, in part, to several factors, including actual increase in student achievement, familiarity with the test forms, differential participation of districts based on their use of the test being normed, and the use of old norms which typically are easier than more recent norms. If student achievement has indeed risen, then out-dated norms would provide a lower standard of comparison than new norms and "thus a state or district whose students score at the current national average would score above the average defined by dated norms" (Linn et al., 1990, p. 9). Linn et al. (1990) also demonstrate that gains in achievement measured by NAEP are much less dramatic than those reported by states and school districts or by test publishers in their norming studies.

The results of a survey of state testing directors suggest that "the conditions for inflated results exist, in some cases to a marked degree" in all states with high-stakes testing programs (Shepard, 1990, p. 20). Shepard (1990, 1991) found that 40 states had high-stakes testing programs that placed "some amount of pressure" (1990, p. 20) on school administrators and teachers to raise test scores. The survey results confirmed that test-curriculum alignment and "teaching the test," rather than test objectives, exist, to unknown degrees, in all of these states. She concludes that "each of these factors will affect the validity of local achievement trends and will also distort the meaning of annual user norms" (1990, p. 20).

Unfortunately, there have been no empirical studies to date to either support or reject the hypothesized lack of motivation generated by the NAEP testing environment, or to show whether students' performance would be improved if motivation were increased. The present study is one of a series of studies being conducted by the NAEP Technical Review Panel with the support of the National Center for Education Statistics to investigate the possibility that NAEP underestimates student achievement due to the low-stakes administration conditions. In this study, NAEP items were administered as a section of a state assessment program that has higher stakes for teachers and principals than those associated with NAEP. Complementary studies involving experimental manipulations of administration conditions are being conducted under the direction of Harold F. O'Neil, Jr., National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (USC subcontract).

Review of the Literature

> . . . ability and motivation have been utterly confounded in test performance and achievement since Binet invented the intelligence test . . . (Atkinson, 1980, p. 18)

Educators and researchers generally agree that motivation is an important factor in school performance and academic achievement (cf. Atkinson & Feather, 1966; Bishop, 1989; Brophy, 1983; Brown & Walberg, in press; Fyans, 1980; Fyans & Maehr, 1990; Igoe, 1991; Matthews & Odom, 1991; Uguroglu & Walberg, 1979). Although much has been written on achievement motivation per se, there has been surprisingly little empirical research on the effects of different motivation conditions on test performance. Before examining the paucity of research on the relationship of motivation and test performance, we first review briefly the general literature on the relationship of motivation and achievement.

Previous Research on Motivation and Achievement

Prior to 1980, achievement motivation theory focused primarily on the need for achievement (*n* achievement) and the effects of test anxiety on test performance (Atkinson & Feather, 1966; Atkinson & Litwin, 1966; Hill, 1980; McClelland, 1955; McClelland, Atkinson, Clark, & Lowell, 1953). Atkinson

and Litwin, in 1966, noted that an increasing number of studies "amply demonstrate that knowledge of motivational differences enhances prediction of achievement related performances" (p. 75). Atkinson (1966, 1980) argues that motivation influences behavior in two primary ways. First, the strength of motivation controls the amount of time an individual devotes to an activity; second, the strength of motivation determines the individual's efficiency in performing that activity. For example, in a situation of "constrained performance," in which there is no opportunity for an individual to select which task to perform (e.g., as in a testing situation) the individual must decide whether to perform the task or leave the situation. If the individual perceives that the consequences of not performing the task are more negative than performing the task, *the level or quality of that performance* then becomes the question of interest (Atkinson, 1966).

Some researchers have hypothesized that the increased use of standardized tests that have no impact on grades has desensitized students so that they care little about their performance, do not try very hard, and do not perform at their ability levels (cf., Brown & Walberg, in press; Paris, Lawton, & Turner, 1992). Others have examined the role that negative or "debilitating" motivational dynamics, such as test anxiety, play in test performance. Hill (1980) argues that at high levels of test anxiety many individuals will perform "well below their optimal level of functioning in the test situation, thereby invalidating their results if one is interested in what the children have learned, as opposed to whether they can demonstrate that learning under heavy testing pressure" (p. 37). Davies (1986), in order to develop a set of "guidelines" for maximizing examination performance, examined motivational variables related to development of ability and acquisition of knowledge. He found that the intensity of motivation (i.e., level of arousal) affects performance on competitive examinations and that, in general, *individuals perform best at intermediate or moderate levels of arousal, while they do least well at the two extremes of high and low levels of arousal.*

Any of the scenarios discussed above could result in poorer performance on tests by students and could explain the relatively poor performance of American students on international tests of achievement and on the NAEP. Regardless of their particular perspective on the issue of the effects of motivation on achievement, however, most educational researchers agree that

motivation influences achievement. For example, the results of a causal analysis by Fyans and Maehr (1990) suggest that student motivation is a "critical mediating variable" in school achievement.[2] Further, the contribution of motivation appears independent of family background and, in the case of sixth, eighth, and tenth graders, outweighs this factor, which historically has been regarded as having the most significant impact on educational outcomes.

Much of the recent discussion on "the crisis in education" focuses on the role of motivation and incentives, whether intrinsic or extrinsic, on student achievement and test performance. Matthews and Odom (1991), for example, examined the relationship between intrinsic motivation and academic achievement among eighth graders in a laboratory school and found that intrinsic motivation for reading and general school orientation[3] were significantly related to all achievement areas measured by the Comprehensive Tests of Basic Skills. The authors conclude that the study findings "clarify the importance of interests, attitudes, wants, desires, and motives in the learning process. With from 4 to 47 percent of the variance in [achievement] being attributable to intrinsic motivation, more emphasis needs to be placed on motivation in the school program" (pp. 39–40).

Rosenbaum (1989) argues that reform efforts to improve student achievement that involve longer school days, longer school years, and higher standards for teachers, curricula, and promotion or graduation focus "too narrowly on symptoms while ignoring the motivation problems that cause poor achievement" (p. 12). At a recent conference on student motivation sponsored by the U.S. Department of Education's Office of Programs for the Improvement of Practice, B. Bradford Brown argued that:

> . . . as schools are presently organized, there "isn't much reason for students to really apply themselves. They know that if they just hang in there, they're going to get a diploma, and they think that's all that counts. . . ."

---

[2] The authors also tested the alternative model, that is, that achievement causes motivation, but their data did not support that causal sequence.

[3] The *Children's Academic Motivation Inventory* by Gottfried (1986) was used to measure intrinsic motivation.

> But school reorganization alone won't be enough. "If we change the schools and forget about the students, the chances for some of the improvements taking hold are very limited." (Mercer, 1990, p. 33)

Several recent studies of motivation and achievement make international comparisons and focus on incentives and labor market consequences of high school achievement (cf. Bishop, 1989; Rosenbaum, 1989). Bishop (1989) concludes that apathy is "the proximate cause of the learning deficit" demonstrated by American high school students when their performances on international achievement tests are compared to those of their international counterparts (p. 3). Bishop argues further that the "fundamental cause" of student apathy is lack of rewards for student effort and learning (p. 6) and that the key to student motivation is the recognition and rewarding of academic effort. In other words, Bishop maintains that if there were perceived consequences for academic effort, students would be more motivated and, by implication, that students who are more motivated would learn more and perform better than those who are less motivated. This conclusion is supported by Gottfried's (1985) findings that, when compared with other elementary and junior high school students, those who demonstrate greater intrinsic academic motivation have better grades, perform better on standardized achievement tests, and demonstrate lower school anxiety.

Two characteristics that have often been used as indicators of student motivation are time spent on learning activities and intensity of student involvement in the learning process. Bishop (1989) notes that the intensity of student involvement is even more important than time devoted to learning, while Goodlad (1983) paints "a general picture of considerable passivity among students . . . " (p. 113). In his survey of high school teachers, Goodlad found that the teachers ranked "lack of student interest" and "lack of parental interest" as the two most important problems facing education. Although Goodlad's survey was conducted a decade ago, his conclusions appear to be just as appropriate today (cf. Bishop, 1989).

Recent Research on Motivation and Test Performance

Despite continuing concern regarding the effects of motivation on student achievement and test performance in general, as well as increasing concern of many education reformers and the National Assessment Governing Board

regarding American students' relatively poor performance on the NAEP, there has been very little empirical research on students' self-reported motivation levels or experimental manipulation of motivational conditions—until recently. Paris et al. (1992) designed several surveys of student attitudes and test-taking strategies which were administered to almost 1000 students in grades 2–11 in Arizona, California, Florida, and Michigan. They found significant age differences in students' perceptions of standardized tests. For example, older students, as opposed to younger students: (a) appeared to be more skeptical regarding the validity of test scores; (b) felt that they were not well informed about the purposes and uses of achievement tests; (c) were concerned that their relative performance on achievement tests would become the basis for comparative social judgment; (d) admitted that they felt ill-prepared to take the tests; (e) reported a lack of good test-taking strategies; and (f) reported less motivation to excel on standardized tests. With respect to the last finding, older students were less likely than younger students to agree with statements such as: "I gave my best effort on the test we took" and "I want to do well on the test because my teacher really cares how well I do." The authors conclude that "whatever the reason, lowered motivation threatens the validity of the test scores" (p. 226).

In another Paris et al. survey (1992) of 250 fourth-, seventh-, and tenth-grade students who took the Michigan Educational Assessment Program Reading Test (a state-mandated, criterion-referenced achievement test), in contrast, most students reported that they tried hard; thought they did well; thought the test was not difficult or confusing; and felt that there was little cheating. However, when the data were examined by age, the results confirmed those of the earlier surveys. Older students, as compared to younger students, cared less about how well they did on the test; thought that parents and teachers did not care about their scores; felt less prepared for the tests; received little explanation and encouragement from their teachers; were more bored by the reading passages and often did not read the entire passage; thought it was "OK" to cheat; reported poor test-taking strategies such as filling in the bubbles without thinking or not going back to check their answers (Paris et al., 1992). An earlier study by Karmos and Karmos (1984) also found that substantial proportions of students in grades 6–9 (n=360) were disaffected by the achievement test process: 47% thought taking achievement tests was a

waste of time; 36% though achievement tests were "dumb"; 30% wanted to get the achievement test over with more than they wanted to perform well; 22% saw no good reason to try to do well; and 21% reported that they do not try very hard on achievement tests.

Although surveys of attitudes toward test-taking, studies of self-reported motivation, and correlational studies of motivation and achievement measures provide useful information, they do not directly measure the effect of motivational conditions in the test situation on achievement test performance. A 1962 study by Burt and Williams reported by Guilford (1967) examined the differences in means on test scores when tests were taken for purposes of promotion or for experimental purposes and the students were aware of the reasons for testing. Burt and Williams found that children who took the test for the purpose of promotion scored, on average, from 3 to 5 points higher than those who took the test for experimental purposes. Variances and reliabilities also were reported to be somewhat higher in the promotion situation. The authors also found that mean scores of students who took examinations to obtain teacher's certificates were 5.8 points higher than those of students who took the examinations for experimental purposes.

Brown and Walberg (in press) have recently completed a study that examines the effects of experimentally-manipulated motivational conditions in the testing environment on the mathematics scores of elementary students in Chicago. Two heterogeneous classes within each of three schools were sampled from grades 3, 4, 6, 7, and 8; the pair of classes selected at each grade level were assigned randomly to the experimental or to the control condition. The Mathematics Concepts subtest (Form 7) of the 1978 Iowa Test of Basic Skills (ITBS) was used to measure mathematics achievement. Teachers of the classes in the control group read the standard instructions for administration of the ITBS, while teachers in the experimental group were instructed to read an additional script emphasizing the importance of the students doing as well as possible for "yourself, your parents, and me" [the teacher] and because the scores would "be compared to students in other grades here at this school as well as to those in other schools in Chicago."

An analysis of variance demonstrated a highly significant main effect of the experimental condition ($F = 10.59$, $p < .01$). The mean Normal Curve Equivalent test score of the students who were asked to "try especially hard"

(the experimental condition) was 41.37, as compared with 36.25 for students given the standard ITBS instructions (the control condition). The effect attributed to motivation is 0.303 standard deviations, which implies that the "average," or typical, test score was raised from the 50th to the 62nd percentile by the additional instructions. The motivational effect was the same for male and female students and across grade levels. The authors concluded that:

> . . . standardized commercial and state-constructed tests which have no bearing on students' grades may be underestimating U.S. students' real knowledge, understanding, skills and other aspects of achievement. To the extent that motivation varies from school to school, moreover, achievement levels of some schools are considerably more underestimated than in others. Such motivational differences would tend to diminish the validity of comparisons of schools and districts. (Brown & Walberg, in press)

NAEP Motivation Studies

As discussed earlier, the issue of the effect of motivation on students' test performance on the NAEP and the validity of its estimates as the "Nation's Report Card" is one of great concern to the National Assessment Governing Board and other education researchers. Recently, several studies designed to assess the existence and potential impact of motivational factors on performance on the NAEP have been completed or currently are in progress. These studies include: (a) the addition of several questions related to issues of cooperation and motivation to the 1991 field test for the 1992 NAEP administration; (b) focus groups composed of twelfth-grade students who participated in the 1991 field test in four states and the District of Columbia; and (c) a multi-component motivation study currently underway.

Field test "motivation" questions. When asked about the difficulty of the mathematics test, half of the eighth-grade students who participated in the field test thought that the math test was "hard" or "very hard" as compared to the other math tests and assignments that they had had that year in school. Recognizing that the NAEP is a low-stakes test for them, 36% of eighth-graders said it was only "somewhat important" or "not very important" for them to do well on the math test. Twenty-eight percent of these students reported that they tried "not at all hard" or "somewhat hard" on the test. Students also were asked "This year in school, how often have you taken mathematics tests where

you were asked to provide detailed solutions to problems you have not worked on before?". Forty-two percent replied "never" and 23% replied "once or twice this year" (Educational Testing Service, 1991). Note that the field test was conducted in February, 1991 when the school year was two-thirds over.

Field test focus groups. Shortly after the field test administration sessions were concluded, the NAEP Field Test administration staff conducted focus groups with subsamples of participating twelfth-grade students. In general, these seniors felt that the assessment allowed them to show how well they could do and that the math test was very difficult. With regard to motivation, most students said that to increase participation:

1. students should not be informed about the test ahead of time;

2. students should not be told that the test does not "count toward their grade";

3. students should be provided with school-level scores so that they could compare themselves to rival schools in their district;

4. extrinsic motivators such as being given the calculator after the session, snacks before and after the test, assemblies, parties, raffles, pizza, etc., would all work to some extent (primarily to increase participation), but such encouragements would not be enough to convince students to do their best. Most students agreed that "motivation came from within and that there was little we could do to motivate a student to do his best on our assessment" because that is a matter of self-pride (Educational Testing Service, 1991, Appendix F); and

5. students probably would take the assessment more seriously and work harder for the NAEP administration staff than for their school staff because the distance traveled by the administration staff makes the assessment seem more important.

NAEP motivation studies: Preliminary results. A multiple-component study to assess the impact of test conditions believed to influence students' motivation on performance on the NAEP currently is being conducted by the NAEP Technical Review Panel researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), housed at UCLA with subcontractors at the University of Southern California and the University of Colorado at Boulder. Prior to an experimental study, in which administration conditions thought to affect student motivation will be experimentally manipulated, researchers at CRESST/UCLA and

CRESST/USC under the direction of Harold F. O'Neil, Jr. conducted (a) focus groups to determine extrinsic rewards that might best motivate students to try their hardest on standardized tests and (b) pilot studies for the main study, which will test the effect of several different tangible and intangible rewards on test performance.

In the focus groups, eighth- and twelfth-graders "brainstormed" to produce a list of incentives in five incentive areas[4] that would most motivate them to work harder on a standardized test. In addition, the participants were asked to select any "incentives" from the list that would discourage them from doing their best (i.e., disincentives). Finally, the students were asked to select the *one* incentive among those in *all* categories that would most motivate them to try their best on a standardized test and also to select the one that would most discourage them. The students overwhelmingly selected a college scholarship as the one incentive (across all categories) that would most motivate them to do their best on a standardized test. A monetary reward was the second most frequently given response. When asked about the greatest disincentive, most students responded "nothing" or gave no response. In the "material reward" category, the most popular motivator was a class party (which was significantly more popular with eighth graders than with twelfth graders). Appearance of a student on television, followed by a letter of recognition to parents, were the two most frequently selected incentives in the "recognition" category. Comparisons of school and individual scores to those of other schools and individuals, respectively, were the considered to be the most encouraging incentives in the "comparisons" category. In the "consequences" category, letting the test score "count" towards regular class grade was the most popular incentive. And, finally, receiving feedback on one's own strengths and weaknesses was the most frequently cited incentive in the "feedback" category. In general, the most frequently selected incentives in each category varied by gender, ethnicity, socio-economic status, and/or grade.

Preliminary results from the pilot studies of the experimental manipulation of test conditions (i.e., tangible reward, intangible reward, and control) indicate no significant differences among the experimental and control groups. That is, the data suggest that groups who received money (the

[4] The incentive areas studied were: material rewards, recognition, comparisons to other groups, consequences, and feedback.

tangible reward condition) or task, ego, or "Walberg-like" instructions[5] (the intangible reward conditions) performed similarly to the control group who received the standard NAEP instructions. However, the data also indicate that prior administration in a school may dilute the results (i.e., "diffusion of treatment" effect). In schools where no prior administration had occurred, eighth-grade students in the experimental conditions scored significantly higher than those in the control group. Further, in the case of eighth graders, ego instructions appeared to have a larger effect on test performance than the other types of instruction or monetary rewards. No significant differences were found for the twelfth-grade students.

A third component of the NAEP motivation study examines the effect of testing environment on NAEP scores. This portion of the study (known as the "State-Embedded Project") is discussed in the remainder of this report.

Methodology

The focus of this study is to investigate the effects of testing environment on student performance on a set of standardized mathematics items. The testing conditions under study are the "low-stakes" environment of the current NAEP administration and a higher stakes environment typified by many state assessment programs. In order to investigate the effects of testing environment, items in the released Block 7 of the NAEP eighth-grade mathematics assessment were embedded in Georgia's state assessment, the Georgia Curriculum-Based Assessments (CBA). The items selected for this study had to meet two conditions. First, they had to have been previously released so that the security of the NAEP item pool could be protected. Second, they had to have been administered to Georgia students in a regular NAEP administration to provide a benchmark against which the state-embedded results could be compared. Block 7 of the mathematics assessment was selected because it met both of these conditions. It was one of the blocks used in the 1990 Trial State Assessment in which Georgia participated, and the block was released following that administration.

---

[5] Task instructions emphasized the intellectual challenge of the task. Ego instructions focused on comparison of an individual's performance with that of others. "Walberg-like" instructions were the same as those used by Brown and Walberg (in press) discussed earlier.

Comparative analyses of student performance on the Block 7 items in the 1990 NAEP Georgia Trial State Assessment and in the 1992 Georgia Curriculum-Based Assessments were conducted to determine whether performance is enhanced when NAEP items are administered in the higher-stakes environment of a regular state assessment that has greater consequences for teachers and schools than does NAEP. Significant differences in performance would suggest that testing environment, and hence, motivation, has an impact on NAEP's estimates of student achievement.

Georgia's assessment was administered to all eighth-grade students and is described in a later section of this report. The Block 7 items of the eighth-grade NAEP mathematics assessment are described below.

The NAEP Block 7 Mathematics Items

Block 7 is a timed administration and contains 18 items administered in 15 minutes. Seventeen items are multiple-choice items and one is an open-ended question. The open-ended format could not be accommodated on Georgia's answer forms; therefore, that item was eliminated from the assessment. In order to lessen the burden on individual examinees, the block was split into two segments, items 1–9 and items 10–16 and 18 (item 17 is the open-ended question). Each segment was administered to approximately 40,000 eighth-grade students in 7-1/2 minutes.

The eighth-grade NAEP items embedded in the Georgia state assessment cover the content areas Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. The Numbers and Operations content area focuses on understanding of numbers, including whole numbers, fractions, decimals, and integers; application to "real-world" situations; and computation and estimation. The Measurement area assesses students' ability to use numbers to describe "real-world" objects. The Geometry content area focuses on students' knowledge of geometric figures and relationships. The Data Analysis, Statistics, and Probability content area focuses on representation and analysis of data across various disciplines; methods for gathering and analyzing data; interpretation of data; and evaluation of arguments based on data analysis. The Algebra and Functions content area for the eighth grade focuses on general algebraic and

functional concepts in an "exploratory" manner. Algebraic expressions at this level may be monomial, polynomial, or rational; may involve more than one variable; and may include mathematical symbols such as those for exponents, radicals, and absolute values (Educational Testing Service, 1988).

Mathematics content of the NAEP Block 7. Classification of each of the 17 NAEP items embedded in the Georgia assessment by content area is presented below.

Content Areas Assessed by Mathematics Block 7 in the Eighth Grade NAEP

Content Area	Item No.	NAEP No.
Numbers & Operations	2	015501
	6	015901
	12	016501
Measurement	1	015401
	4	015701
	9	016201
Geometry	3	015601
	10	016301
	11	016401
	13	016601
	14	016701
Data Analysis, Statistics, & Probability	5	015801
	8	016101
	17[6]	017001
Algebra & Functions	7	016001
	15	016801
	16	016901

A copy of the first nine mathematics items of Block 7 included in Forms 1 and 4 of the 1992 Georgia Curriculum-Based Assessments is provided in Appendix A. The last eight items of Block 7 included in Forms 2 and 3 of the 1992 CBA are listed in Appendix B.

[6] The item referred to here as item 17 is the original item 18 in Block 7 (the original item 17 was omitted because of the open-ended format). For the sake of simplicity and continuity, original item 18 (017001) will be referred to as item 17 in the remainder of this report.

The 1992 Georgia Curriculum-Based Assessments

In 1971 the state of Georgia established a statewide testing program to aid in instructional planning, evaluation of educational programs, and provision of feedback to local systems, schools, and students. The assessment program has been revised and expanded twice since 1971, first in 1986 and again in 1992. The Spring 1992 assessment is the first administration of the new testing program, the Georgia Curriculum-Based Assessments (CBA). The CBA was administered during May 1992 to all third-, fifth-, and eighth-grade students in Georgia (i.e., to all those present in school on their assigned or make-up testing days). The stated purpose of the assessments is to measure students' knowledge and achievement for a broad range of the state's curriculum; to "answer the question 'How well are students learning the state-required objectives of the Quality Core Curriculum?' " (Rogers 1992c); and to assess higher-order thinking skills as well as specific, factual knowledge. The content areas assessed are mathematics, reading, science, and social studies. In addition, health items are included in the assessments of fifth- and eighth-grade students.

Uses of the Georgia CBA. According to Werner Rogers (1992b), the Georgia State Superintendent of Schools, "CBA results will be used to guide state policy decisions, to support and encourage state curriculum goals, and to hold the state accountable for its constitutional responsibility to provide a thorough and efficient system of public education" (p. 2). At the state level, administrators use the assessment results to monitor student achievement statewide; to allocate funding for remediation programs based on student need; to assist local school systems in evaluating their needs and in planning and implementing programs for improving their curriculum and instruction; and to develop policies regarding curriculum, instruction, and administration. Local administrators use the test results to evaluate their curricular and instructional programs and to identify strengths and weaknesses. Test results also are used in setting local priorities for resource allocation, staffing and staff development, and instruction and instructional materials (Rogers, 1992c).

Students do not receive their individual scores, and therefore, the CBA might be considered "low-stakes" for students; however, test results *are* of consequence to local systems. Because the results have potential consequences

for school administrators and teachers, it is expected that teachers will devote greater effort to encouraging students to put forth their best efforts on the Georgia assessment than on NAEP, which is not even reported below the state level. The primary direct appeal to students' motivation in the standard administration directions is at the ego level. The general instructions to students at the beginning of testing are as follows:

> Now we are going to take a test to see *how well you can do* [emphasis added] in such things as language arts-reading, mathematics, social studies, science and health. The test contains exercises similar to those we have in class every day. (Rogers, 1992a, p. 9)

Matrix sampling procedures. The CBA uses a form of matrix sampling with a spiraling design in which different students are tested on different items in each content area. In this case, four parallel test forms are administered at each grade level in each school. Not only are the items included in each content area different on each form, but the order of presentation of each content area also varies by form. The test books for each grade in each school are pre-packaged so that when they are distributed in the order in which they are packaged each grade is stratified into four subgroups by test form.

Mathematics content of the CBA. The content areas included in the mathematics section of the CBA are:

- Numbers and Number Relations;
- Operations and Computations;
- Geometry;
- Measurement; and
- Probability and Statistics.[7]

In the 1992 administration at grade 8, the NAEP mathematics items were included as the last section, Section 6, in each of the four forms of the CBA. The first nine items from the NAEP mathematics Block 7 were included in Test Forms 1 and 4, while the remaining eight items in Block 7 were included

[7] The CBA content areas are similar to the mathematics content areas assessed in Block 7 of the NAEP: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions.

in Forms 2 and 3. Since the CBA uses a spiraling block design, the NAEP items were immediately preceded by different content areas in each test form. The first nine NAEP items followed the content areas Language Arts: Reading in Form 1 and Social Studies in Form 4. In Forms 2 and 3, the last eight NAEP items followed the sections on Mathematics and Science, respectively.

Administration procedures. The CBA was administered during the period May 4–22, 1992 to all regular third-, fifth-, and eighth-grade students[8] present in school on the testing days. The testing was conducted by teachers in the school, and all testing rooms were monitored by a School Test Coordinator. The assessment for the eighth grade requires a total of about 4-1/2 hours. The state Department of Education (personal communication) recommends that testing be conducted during either a two- or three-day period within the May 4-22 testing period and that students be tested on three (or two) sections per day (six sections total). However, the Department does not mandate the scheduling or the number of days for testing. Local systems determine their own scheduling within the May 4-22 period and may, in fact, choose to administer only one section per day or to administer the entire test in one day. However, most systems administered the CBA on two days with three sections per day. Make-up sessions were scheduled for students absent during their assigned testing period.

All materials, the Examiner's Manuals, test books, scratch paper, used answer sheets, and all unused materials were returned to the School Testing Coordinator. The used answer sheets and scratch paper were returned to Test Scoring and Reporting Services of the University of Georgia for scoring, while all test books, manuals, and unused answer sheets were destroyed by the school.

Data analysis. The data file containing the item responses from the 1992 administration of the NAEP as part of the Georgia CBA was provided by the University of Georgia Test Scoring and Reporting Services. The file contains a total of 80,836 useable student records. A total of 40,403 students (20,214 for Form 1 and 20,189 for Form 4) responded to the first nine items from Block 7 of

[8] Students with disabilities or limited English proficiency (LEP) who participate in regular instructional programs may be included in the assessment *if* the local school system desires. However, special codes are provided so that data for officially classified LEP students or students with disabilities may be removed from school and system reports of summary results.

NAEP. A nearly identical number of students, 40,433, responded to the last eight NAEP items from Block 7, with 20,631 responding to Form 2 and 19,802 to Form 3.

Statistics (e.g., means, standard deviations, standard errors, frequency distributions, and item statistics) were computed separately for the first nine and the last eight NAEP items. These descriptive statistics were computed for the two combined forms including the same items (i.e., Forms 1 and 4; Forms 2 and 3), as well as for individual forms. They also were computed for the total sample and for male, female, white, and black subpopulations.

The results for Block 7 items when administered in 1990 as part of the Trial State Assessment (TSA) provided the benchmark against which the results for the 1992 administration of the same items embedded in the Georgia CBA were compared. Weighted estimates of the same parameters estimated for the 1992 data collection were obtained for the 1990 TSA data.

Results

Total Scores

The means, their standard errors, and .95 confidence intervals, as well as standard deviations, are listed in Table 1 for the two subsets of items from Block 7 of the NAEP grade 8 mathematics assessment (see Appendix C for all Tables and Figures). Results from the administration of these items to Georgia students as part of the 1990 Trial State Assessment (TSA) are listed on the left and results from the administration of these items as part six of the 1992 Georgia Curriculum-Based Assessments (State Embedded Results) are listed on the right in Table 1. As can be seen, the mean for the first nine items was significantly higher in the 1992 state-embedded administration (i.e., in the Georgia CBA) than in the 1990 TSA administration. The means for the last eight items (10 through 17) on the two administrations are not significantly different, however. Furthermore the sum of the 1992 means for the two subsets of items, while higher than the 1990 TSA mean, falls within the .95 confidence interval for the 1990 TSA mean.

The significant difference on the first nine items represents an effect size (difference in means divided by the 1992 standard deviation) of .18. This effect size is close to the .2 that Cohen (1988) proposed as a "small" effect size for the

difference between means. Although not significant, the effect size for the last eight items is in the opposite direction (-.04). Additional analyses were conducted to evaluate the sensitivity of these overall results to differences in omit and non-response rates on the TSA and State Embedded administrations. Since those analyses merely confirmed the significant positive effect size for the first nine items and the essentially zero effect size for the last eight items in the Georgia CBA, the details of those results are not presented here.

Histograms comparing the distributions of the 1990 TSA and 1992 State Embedded administrations are shown in Figures 1 and 2 for the first nine and last eight items, respectively (see Appendix C for all Tables and Figures. With the exception of the somewhat larger percentage of scores of zero in 1992 than in 1990, the shift to the right in the 1992 distribution in comparison to the 1990 distribution is consistent with the increase in the mean (higher percentages of scores of 6 and above and lower percentages of scores of 1 through 5 in 1992 than 1990). With the possible exception of the larger number of scores of zero in 1992 than 1990, the distributions for the last eight items for the two years shown in Figure 2 are very similar. As would be expected by the low means, the distributions have a noticeable positive skew (.770 and .820 in 1990 and 1992, respectively).

Results parallel to those shown in Table 1 for the total samples of students in the two years are shown separately for males and females in Table 2. The significant mean differences between 1990 and 1992 obtained for the first nine items for the total samples were replicated for both males and females. The effect sizes for males and females were nearly identical (.17 for males and .18 for females). For the last eight items, the differences in 1990 and 1992 were not significant for either males (effect size = -.01) or females (effect size = -.08).

As can be seen in Table 3, the pattern of significant difference for the first nine items and no significant difference for the last eight items holds when the results for white and black students were analyzed separately. The effect sizes for the first nine items were .17 and .24 for white and black students, respectively. The corresponding effect sizes for the last eight items were -.05 and -.01. The sample sizes of other racial/ethnic groups taking Block 7 in the 1990 TSA were too small for meaningful comparisons.

Item Level Results

A scatterplot of the state average percent correct values for the 1992 state-embedded administration with the weighted average percent correct values from the 1990 TSA administration is shown in Figure 3 for the 17 items in both administrations. There is a strong relationship between the item percent correct values for the two administrations. The correlation between the two sets of percent correct values was .97. As can be more readily seen in Figure 4, the difference in percent correct from 1990 to 1992 was less than 10 in absolute value for all the items. The difference in percent correct was less than 5 in absolute value for nine of the items; the 1992 percent correct was higher than that in 1990 by 5 to 7 percent for seven items; and the percent correct was lower by 8 percent in 1992 than 1990 on the remaining item.

Tables 4 through 8 list the percent correct values for individual items arranged by the five NAEP content categories: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. In each of the tables the percent correct values in 1990 and in 1992 are reported for the total samples and separately for male, female, white, and black students.

For the three Numbers and Operations items, the 1992 percent correct values for the total sample exceeded the corresponding 1990 values by between 2.7 and 7.0 percent (see Table 4). The first two of these items, which have the larger differences, were among the first nine items while the third item was among the last eight items. The corresponding differences for the three Measurement items (Table 5), which were all included in the first nine items, range from -3.3 to 5.8 percent.

One of the five Geometry items (Table 6, second item) had the largest decline in percent correct (-8.2) between 1990 and 1992. That item was the first item in the subset of the last eight items when administered in the 1992 embedded assessment. Two of the other Geometry items, one appearing among the first nine and one among the last eight, had increases of 5.6 and 5.7 percent, respectively.

The first two Data Analysis, Statistics, and Probability items (Table 7) were among the first nine and had increases in percent correct from 1990 to 1992 of 7.0 and 4.8 percent. The third of these items was among the last eight

and had a decrease of 4.8 percent. In a similar pattern, the only Algebra and Functions item showing a substantial increase (7.0%) from 1990 to 1992 was among the first nine. The two Algebra and Functions items that were part of the subset of the last eight items had changes in percent correct of .3 and -2.0.

Graphical displays of the percent correct values for 1990 and 1992 are provided in Figures 5 through 21. In each of these figures the percent correct is reported for the total sample and for male, female, white, and black subsamples. Standard errors of the percent correct estimate also are displayed; however, the standard errors are so small for the 1992 results (due to the large sample size) that they are not always visible.

From an inspection of Figures 5 through 21 it is apparent that there is great consistency in changes in percent correct from 1990 to 1992 for the total sample and for each of the four subpopulations. With relatively few exceptions, an item with an increase (or decrease) in percent correct from 1990 to 1992 for one subpopulation had a similar increase (decrease) for all subpopulations.

Form Differences

Since each of the item sets was administered in two of the Georgia test forms, it is possible to have some check on the effects of contextual factors associated with different forms. The first 9 items were included as section 6 of Forms 1 and 4 and the last 8 items were included as section 6 of Forms 2 and 3. The means, standard deviations, and standard errors of the means for the NAEP item sets are shown by form of the 1992 Georgia test in Table 9. Also listed in Table 9 is the correlation of each NAEP item set with the mathematics section of the Georgia test form. The order of the content in the five operational sections of the four Georgia forms is shown in the footnotes of Table 9.

There are small differences in the means of the NAEP item subsets as a function of test form. Using the standard deviation of the first nine items when administered with Form 1 and the standard deviation of the last eight items when administered with Form 2 as the metric, the effect sizes were .04 and .15 for the first nine and the last eight items, respectively. Since the identical NAEP item sets were administered in Forms 1 and 4 and in Forms 2 and 3 to randomly equivalent samples, the small mean differences are

presumably attributable to context effects provided by the different orders of the content in the first five sections of the forms.

Discussion

This study is one component of a series of studies undertaken to investigate the hypothesis that NAEP results yield an underestimate of U.S. students' achievement levels because students are not motivated to do their best due to a lack of any stakes associated with the results. This study is intended to complement the studies involving experimental manipulations of administration conditions that are being directed by Harold F. O'Neil, Jr. (CRESST/USC) by including NAEP items in a regular administration of a state assessment that has higher stakes than NAEP.

Experience with the 1986 NAEP "reading anomaly" (Beaton & Zwick, 1990; Haertel, 1989) demonstrated that "seemingly minor changes in the administrative procedures, assessment booklets, and the timing" (Beaton, 1990, p. 8) can result in noticeable changes in scores on a NAEP assessment. For example, the mean difference in percent correct for the 23 items administered in both 1984 and 1986 corresponded to an effect size of .12 (1984 average minus 1986 average divided by 1986 standard deviation) (Mislevy, 1990). Although a number of factors, including real differences in performance, may have contributed to that difference, it is generally believed that subtle changes in the context in which items were administered accounted for a large fraction of the "anomaly."

There are certainly contextual differences in the administration of the Block 7 NAEP items as part of the 1992 CBA in comparison to the earlier administration in the 1990 TSA. Practical time constraints, for example, forced the split of the block of 17 multiple-choice items into two subsets of the first 9 and the last 8 items for the State Embedded administrations. Another factor that could affect the results is the difference in the times of testing. The analysis is based not only on the performances of two different cohorts of eighth-grade students (1990 and 1992 cohorts), but also on the performances of students tested in February (1990) and of those tested in May (1992). Thus, small increases may be due to the additional 2–3 months of instruction students in the May administration received. For these reasons, it would be unwise to over-interpret small differences in means between 1990 and 1992.

The major concern of the hypothesis that low motivation leads to an underestimate of student achievement, however, is not with small differences. Rather, the concern is that achievement is being underestimated by a substantial amount because of the lack of stakes. Thus, the potential value of the present study derives from the possibility that it would identify a major change in estimated average achievement either for all students or for a major subpopulation.

Significant differences between performance in the 1992 State Embedded administration and the 1990 TSA administration were found for only one of the two subsets of items. It may be that the difference in results for the two subsets of items is a function of their relative difficulty. Increasing the stakes may have a greater influence on the subset of relatively easier items than on the subset of relatively difficult items, because trying harder can increase performance on material that low and moderate achieving students know how to do but not on more difficult material that they do not know how to do. High achieving students who are capable of answering the harder items correctly, on the other hand, may be less influenced by changing the stakes of the administration. However, it was the easiest of the eight items on the second subset that showed the largest decline (-.08) between the 1990 TSA and the 1992 State Embedded administration. This item was the first item in the second subset of items. The sharp drop on the first item might be the result of the contextual effect of an item appearing in the first position in 1992 rather than as the tenth item in the full block of items in 1990.

It should be recognized that even for the subset of items where the differences were significant, the effect size was relatively small (.18). Comparison of this effect size to the effect size mentioned above (.12) associated with the 1986 reading anomaly at age 17 provides an additional perspective on the magnitude of the difference. That is, the effect size for the first nine items in the present study is only 1.5 times that observed in the reading anomaly that was largely attributed to context effects. The form-to-form differences in means on the same subsets of NAEP items shown in Table 9 (effect sizes of .04 and .15) provide an additional indication that some fluctuation in means is to be expected due to changes in the context of other items administered.

The relatively small difference for only one of the two subsets of items, together with the possibilities that there may have been some effects due to

contextual changes in the administrations and that there may be some real between-year differences in achievement, lead to the conclusion that estimates of achievement from NAEP would not be substantially higher if the stakes were increased to the level associated with the Georgia Curriculum-Based Assessments. It is possible that increasing the stakes still further by adding, for example, rewards and sanctions for individual students, would yield a larger effect. But the relatively high-stakes uses of school-level results for reporting, evaluation, and allocation of remedial funds that are associated with the Georgia program yield results that are not markedly better than those obtained in the administration as part of the Trial State Assessment.

References

Allington, R. L., & McGill-Franzen, A. (1992). Does high-stakes testing improve school effectiveness? *ERS Spectrum, 10,* 3-12.

Atkinson, J. W. (1966). Motivational determinants of risk-taking behavior. In J. W. Atkinson & N. T. Feather (Eds.), *A theory of achievement motivation*. New York: John Wiley & Sons, Inc.

Atkinson, J. W. (1980). Motivational effects in so-called tests of ability and educational achievement. In L. J. Fyans, Jr. (Ed.), *Achievement motivation: Recent trends in theory and research*. New York: Plenum Press.

Atkinson, J. W., & Feather, N. T. (Eds.). (1966). *A theory of achievement motivation*. New York: John Wiley & Sons, Inc.

Atkinson, J. W., & Litwin, G. H. (1966). Achievement motive and test anxiety conceived as motive to approach success and motive to avoid failure. In J. W. Atkinson & N. T. Feather (Eds.), *A theory of achievement motivation*. New York: John Wiley & Sons, Inc.

Beaton, A. E. (1990). Introduction. In A. E. Beaton & R. Zwick (Eds.), *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E., & Zwick, R. (Eds). (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Bishop, J. (1989). *Incentives for learning: Why American high school students compare so poorly to their counterparts overseas* (Working Paper #89-09). Ithaca, NY: Cornell University, Center for Advanced Human Resource Studies and New York State School of Industrial and Labor Relations.

Brophy, J. (1983). Conceptualizing student motivation. *Educational Psychologist, 18,* 200-215.

Brown, S. M., & Walberg, H. J. (in press). *Motivational effects on test scores of elementary-school students: An experimental study.*

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average.* Daniels, WV: Friends for Education.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davies, D. (1986). *Maximizing examination performance: A psychological approach*. London: Kogan Page Ltd.

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

Educational Testing Service. (1991). *The results of the NAEP 1991 field test for the 1992 National and Trial State Assessments*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Fyans, L. J., Jr. (Ed.). (1980). *Achievement motivation: Recent trends in theory and research*. New York: Plenum Press.

Fyans, L. J., Jr., & Maehr, M. L. (1990). *School "culture," motivation, and achievement* (Project Report). Urbana: University of Illinois.

Goodlad, J. (1983) *A place called school*. New York: McGraw-Hill.

Gottfried, A. E. (1985). Academic intrinsic motivation in elementary and junior high school students. *Journal of Educational Psychology, 77*, 631-645.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Haertel, E. (Chair). (1989). *Report of the NAEP Technical Review Panel on the 1986 reading anomaly, the accuracy of NAEP trends, and issues raised by state-level NAEP comparisons* (National Center for Education Statistics Tech. Rep. No. CS 89-499). Washington, DC: U.S. Office of Educational Research and Improvement.

Hill, K. T. (1980). Motivation, evaluation, and educational testing policy. In L. J. Fyans, Jr. (Ed.), *Achievement motivation: Recent trends in theory and research*. New York: Plenum Press.

Igoe, A. R. (1991). Learners and self-reports of characteristics related to academic achievement and motivation. In *Proceedings of selected research presentations at the annual convention of the Association for Educational Communications and Technology*.

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

Linn, R. L., Graue, M.E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "Everyone is above average." *Educational Measurement: Issues and Practice, 9*, 5-14.

Matthews, D. B., & Odom, B. L. (1991). Intrinsic motivation: A major factor in student academic achievement. *NALS Journal, 15*, 32-42.

McClelland, D. C. (1955). *Studies in motivation*. New York: Appleton-Century-Crofts.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. New York: Appleton-Century-Crofts.

Mercer, J. (1990). Conference explores role of student motivation in correcting educational ills. *Black Issues in Higher Education, 7*, 32-33.

Mislevy, R. J. (1990). Item-by-form variation in 1984 and 1986 NAEP reading surveys. In A. E. Beaton & R. Zwick (Eds.), *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Paris, S. G., Lawton, T. A., & Turner, J. C. (1992). Reforming achievement testing to promote students' learning. In C. Collins & J. N. Mangieri (Eds.), *Teaching thinking: An agenda for the twenty-first century*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rogers, W. (1992a). *Curriculum-based assessment: Examiner's manual* (Grade 8). Atlanta: Georgia Department of Education.

Rogers, W. (1992b). *Curriculum-based assessment: Overview and content description*. Atlanta: Georgia Department of Education.

Rogers, W. (1992c). *Georgia Statewide Student Assessment Program* (Pamphlet). Atlanta: Georgia Department of Education.

Rosenbaum, J. E. (1989). What if good jobs depended on good grades? *American Educator, 13*, 10-15.

Shanker, A. (1990, July 29). How much do our kids really know? Raising the stakes on NAEP. *The New York Times*.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*, 5-22.

Shepard, L. A. (1991). The influence of standardized tests on the early childhood curriculum, teachers, and children. In B. Spodek & O. N. Saracho (Eds.), *Yearbook in Early Childhood Education* (Vol. 2). New York: Teachers College Press.

Uguroglu, M. E., & Walberg, H. J. (1979). Motivation and achievement: A quantitative synthesis. *American Educational Research Journal, 16*, 375-389.

APPENDIX A

NAEP Block 7 Mathematics Items Included in Forms 1 and 4 of the Eighth Grade Assessment: 1992 Georgia Curriculum-Based Assessments

1 How many hours are equal to 150 minutes?

(A) $1\frac{1}{2}$

(B) $2\frac{1}{4}$

(C) $2\frac{1}{3}$

(D) $2\frac{1}{2}$

(E) $2\frac{5}{6}$

M015401

2 If $\frac{2}{25} = \frac{n}{500}$, then $n =$

(A) 10

(B) 20

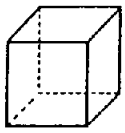
(C) 30

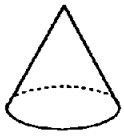
(D) 40

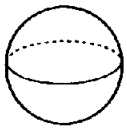
(E) 50

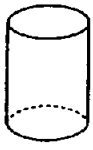
M015501

3 A straight line segment could NOT be drawn on the surface of which of the following solids?

(A) 

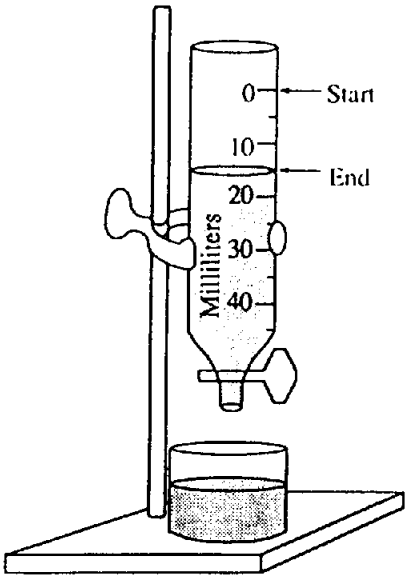
(B) 

(C) 

(D) 

(E) 

M015601



4 In the figure above, the tube was filled to the 0 mark at the start. How much liquid has been let out?

(A) 10 milliliters

(B) 15 milliliters

(C) 25 milliliters

(D) 40 milliliters

(E) 50 milliliters

M015701

5 The average weight of 50 prize-winning tomatoes is 2.36 pounds. What is the combined weight, in pounds, of these 50 tomatoes?

(A) 0.0472

(B) 11.8

(C) 52.36

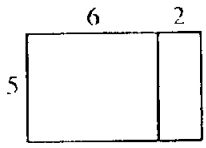
(D) 59

(E) 118

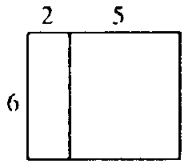
M015801

6 Which of the following figures best illustrates the statement $5 \times (6 + 2) = (5 \times 6) + (5 \times 2)$?

(A)

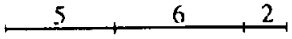


(B)

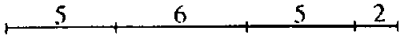


(C)

(D)



(E)



M015901

7 What is the least whole number x for which $2x > 11$?

(A) 5

(B) 6

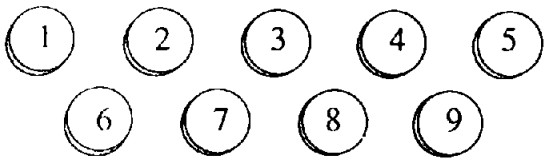
(C) 9

(D) 22

(E) 23

M016001

GO TO NEXT PAGE



8 The nine chips shown above are placed in a sack and then mixed up. Madeline draws one chip from this sack. What is the probability that Madeline draws a chip with an even number?

(A) $\frac{1}{9}$

(B) $\frac{2}{9}$

(C) $\frac{4}{9}$

(D) $\frac{1}{2}$

(E) $\frac{4}{5}$

M016101

9 If a measurement of a rectangular box is given as 48 cubic inches, then the measurement represents the

(A) distance around the top of the box

(B) length of an edge of the box

(C) surface area of the box

(D) volume of the box

M016201

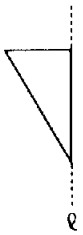
STOP HERE

APPENDIX B

NAEP Block 7 Mathematics Items Included in Forms 2 and 3 of the Eighth Grade Assessment: 1992 Georgia Curriculum-Based Assessments

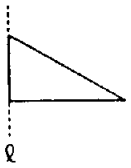
1 Which of the following shows the result of flipping the above triangle over the line ℓ?

(A)

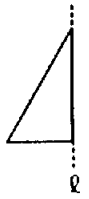


(B)

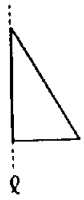
(C)



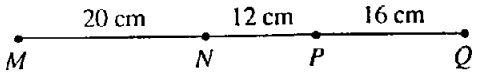
(D)



(E)



M016301



2 What is the distance between the midpoint of MN and the midpoint of PQ shown above?

(A) 18 cm

(B) 24 cm

(C) 26 cm

(D) 28 cm

(E) 30 cm

M016401

3 The least common multiple of 8, 12, and a third number is 120. Which of the following could be the third number?

(A) 15

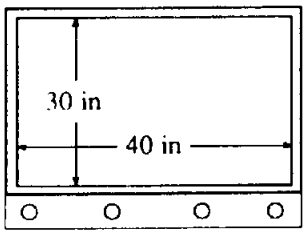
(B) 16

(C) 24

(D) 32

(E) 48

M016501



4 What is the diagonal measurement of the TV screen shown in the figure above?

(A) 25 inches

(B) 35 inches

(C) 50 inches

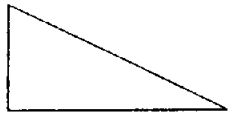
(D) 70 inches

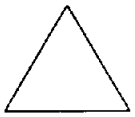
(E) 1,200 inches

M016601

GO TO NEXT PAGE

5 Which of the following figures contains line segments that are perpendicular?

(A) 

(B) 

(C) 

(D) 

M016701

6 The length of a rectangle is 3 more than its width. If L represents the length, what is an expression for the width?

(A) $3 \div L$

(B) $L \div 3$

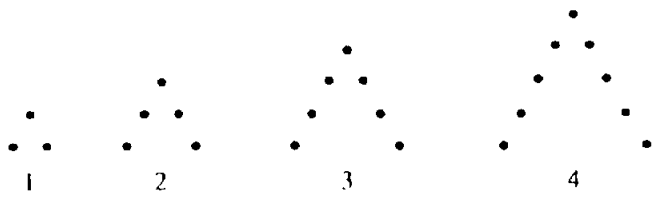
(C) $L \times 3$

(D) $L + 3$

(E) $L - 3$

M016801

GO TO NEXT PAGE



7 If this pattern of dot-figures is continued, how many dots will be in the 100th figure?

(A) 100

(B) 101

(C) 199

(D) 200

(E) 201

M016901

8 There are 15 girls and 11 boys in a mathematics class. If a student is selected at random to run an errand, what is the probability that a boy will be selected?

(A) $\frac{4}{26}$

(B) $\frac{11}{26}$

(C) $\frac{15}{26}$

(D) $\frac{11}{15}$

(E) $\frac{15}{11}$

M017001

STOP HERE

APPENDIX C

Tables and Figures

Table 1

Means, Standard Deviations, Standard Errors of the Means, and .95 Confidence Intervals for the 1990 Trial State Assessment (TSA) and the 1992 State Embedded Administrations

Statistic	1990 TSA results	1992 State embedded results
Sum of Items 1 through 9 of Block 7		
Mean	4.837	5.238
Standard deviation	2.162	2.280
Standard error	0.126	46
46	4.589 to 5.085	5.216 to 5.260
Sum of Items 10 through 17 of Block 7		
Mean	2.429	2.357
Standard deviation	1.647	1.700
Standard error	0.096	0.008
.95 Confidence intervals for means	2.240 to 2.618	2.341 to 2.373
Sum of Items 1 through 17 of Block 7		
Mean	7.266	7.595
Standard deviation	3.360	NA
Standard error	0.196	NA
.95 Confidence interval for TSA mean	6.880 to 7.652	NA

Note. NA: Not available due to split of Block 7 into two sub-blocks of 9 and 8 items each.

Table 2

Means, Standard Deviations, Standard Errors of the Means, and .95 Confidence Intervals by Gender for the 1990 Trial State Assessment (TSA) and the 1992 State Embedded Administrations

	1990 TSA results		1992 State results	
Statistic	Male	Female	Male	Female
Sum of Items 1 through 9 of Block 7				
Mean	4.785	4.893	5.175	5.306
Standard deviation	2.182	2.147	2.316	2.239
Standard error	0.178	0.180	0.016	0.016
.95 Confidence intervals for means	4.433 to 5.137	4.537 to 5.249	5.144 to 5.206	5.275 to 5.337
Sum of Items 10 through 17 of Block 7				
Mean	2.380	2.481	2.367	2.349
Standard deviation	1.676	1.619	1.745	1.654
Standard error	0.137	0.136	0.012	0.012
.95 Confidence intervals for means	2.109 to 2.651	2.212 to 2.750	2.343 to 2.391	2.325 to 2.373
Sum of Items 1 through 17 of Block 7				
Mean	7.165	7.374	7.542	7.655
Standard deviation	3.418	3.307	NA	NA
Standard error	0.278	0.278	NA	NA
.95 Confidence interval for TSA mean	6.615 to 7.715	6.824 to 7.924	NA	NA

Note. NA: Not available due to split of Block 7 into two sub-blocks of 9 and 8 items each.

Table 3

Means, Standard Deviations, Standard Errors of the Means, and .95 Confidence Intervals by Racial Group for the 1990 Trial State Assessment (TSA) and the 1992 State Embedded Administrations

	1990 TSA results		1992 State results	
Statistic	White	Black	White	Black
Sum of Items 1 through 9 of Block 7				
Mean	5.367	3.857	5.731	4.370
Standard deviation	2.056	2.020	2.188	2.146
Standard error	0.150	0.201	0.014	0.018
.95 Confidence intervals for means	5.071 to 5.663	3.458 to 4.256	5.704 to 5.758	4.335 to 4.405
Sum of Items 10 through 17 of Block 7				
Mean	2.758	1.804	2.662	1.793
Standard deviation	1.719	1.273	1.792	1.322
Standard error	0.126	0.127	0.011	0.011
.95 Confidence intervals for means	2.509 to 3.007	1.552 to 2.056	2.640 to 2.684	1.771 to 1.815
Sum of Items 1 through 17 of Block 7				
Mean	8.125	5.661	8.393	6.163
Standard deviation	3.313	2.808	NA	NA
Standard error	0.242	0.279	NA	NA
.95 Confidence interval for TSA mean	7.648 to 8.602	5.107 to 6.215	NA	NA

Note. NA: Not available due to split of Block 7 into two sub-blocks of 9 and 8 items each.

Table 4

State Average Percent Correct on the Three NAEP Numbers and Operations Items by Year and Condition of Administration by Subpopulation (Standard Errors in Parentheses)

NAEP item number	Year	Admin. condition	Subpopulation				
			Total	Male	Female	White	Black
M015501	1990	TSA	46.4	43.9	49.0	52.2	35.2
			(2.9)	(4.1)	(4.2)	(3.7)	(4.8)
	1992	State Assess	51.7	49.3	54.1	56.9	42.0
			(0.2)	(0.4)	(0.4)	(0.3)	(0.4)
M015901	1990	TSA	36.7	32.0	41.8	41.1	29.0
			(2.8)	(3.8)	(4.2)	(3.6)	(4.5)
	1992	State Assess	43.7	40.6	46.8	46.8	37.7
			(0.2)	(0.3)	(0.4)	(0.3)	(0.4)
M016501	1990	TSA	16.9	15.5	18.3	19.8	11.0
			(2.2)	(3.0)	(3.3)	(2.9)	(3.1)
	1992	State Assess	19.6	19.2	20.0	22.5	14.1
			(0.2)	(0.3)	(0.3)	(0.3)	(0.3)

Table 5

State Average Percent Correct on the Three NAEP Measurement Items by Year and Condition of Administration by Subpopulation (Standard Errors in Parentheses)

NAEP item number	Year	Admin. condition	Subpopulation				
			Total	Male	Female	White	Black
M015401	1990	TSA	57.6	59.9	55.1	62.6	47.7
			(2.9)	(4.0)	(4.2)	(3.5)	(5.0)
	1992	State Assess	58.4	58.5	58.4	65.3	46.3
			(0.2)	(0.3)	(0.3)	(0.3)	(0.4)
M015701	1990	TSA	90.2	91.2	89.2	94.3	82.2
			(1.7)	(2.3)	(2.6)	(1.7)	(3.8)
	1992	State Assess	86.9	87.6	86.3	91.6	79.1
			(0.2)	(0.2)	(0.2)	(0.2)	(0.3)
M016201	1990	TSA	39.3	41.0	37.4	45.1	28.2
			(2.9)	(4.0)	(4.1)	(3.6)	(4.5)
	1992	State Assess	45.1	45.4	44.8	47.9	36.5
			(0.2)	(0.4)	(0.4)	(0.3)	(0.4)

Table 6

State Average Percent Correct on the Five NAEP Geometry Items by Year and Condition of Administration by Subpopulation (Standard Errors in Parentheses)

NAEP item number	Year	Admin. condition	Subpopulation				
			Total	Male	Female	White	Black
M015601	1990	TSA	64.7	62.1	67.5	67.8	59.1
			(2.8)	(4.0)	(3.9)	(3.4)	(4.9)
	1992	State Assess	70.4	69.3	71.6	73.3	65.5
			(0.2)	(0.3)	(0.3)	(0.3)	(0.4)
M016301	1990	TSA	66.7	63.7	69.8	71.7	57.3
			(2.8)	(3.9)	(3.9)	(3.3)	(4.9)
	1992	State Assess	58.5	55.4	61.6	63.2	50.5
			(0.2)	(0.3)	(0.3)	(0.3)	(0.4)
M016401	1990	TSA	28.4	28.4	28.5	33.1	19.9
			(2.6)	(3.7)	(3.8)	(3.4)	(4.0)
	1992	State Assess	28.7	29.9	27.5	34.6	18.1
			(0.2)	(0.3)	(0.3)	(0.3)	(0.3)
M016601	1990	TSA	20.5	22.2	18.7	25.7	10.1
			(2.4)	(3.4)	(3.3)	(3.2)	(3.0)
	1992	State Assess	26.1	29.1	23.0	32.6	14.1
			(0.2)	(0.3)	(0.3)	(0.3)	(0.3)
M016701	1990	TSA	22.1	23.2	20.9	24.7	16.9
			(2.4)	(3.4)	(3.4)	(3.2)	(3.7)
	1992	State Assess	21.0	20.8	21.1	23.0	16.8
			(0.2)	(0.3)	(0.3)	(0.3)	(0.3)

Table 7

State Average Percent Correct on the Three NAEP Data Analysis, Statistics, and Probability Items by Year and Condition of Administration by Subpopulation (Standard Errors in Parentheses)

NAEP item number	Year	Admin. condition	Subpopulation				
			Total	Male	Female	White	Black
M015801	1990	TSA	44.3	45.9	42.6	53.4	27.5
			(2.9)	(4.1)	(4.2)	(3.7)	(4.5)
	1992	State Assess	51.3	53.4	49.3	59.4	37.0
			(0.2)	(0.4)	(0.4)	(0.3)	(0.4)
M016101	1990	TSA	64.3	63.2	65.4	73.3	48.5
			(2.8)	(3.9)	(4.0)	(3.2)	(5.0)
	1992	State Assess	69.1	66.7	71.6	75.9	58.0
			(0.2)	(0.3)	(0.3)	(0.3)	(0.4)
M017001	1990	TSA	39.2	32.9	46.0	42.1	34.4
			(2.9)	(3.8)	(4.2)	(3.6)	(4.7)
	1992	State Assess	34.4	31.5	37.4	35.9	31.7
			(0.2)	(0.3)	(0.3)	(0.3)	(0.4)

Table 8

State Average Percent Correct on the Three NAEP Algebra and Functions Items by Year and Condition of Administration by Subpopulation (Standard Errors in Parentheses)

NAEP item number	Year	Admin. condition	Subpopulation				
			Total	Male	Female	White	Black
M016001	1990	TSA	40.3	39.3	41.4	46.9	28.3
			(2.9)	(4.0)	(4.2)	(3.7)	(4.5)
	1992	State Assess	47.3	46.9	47.7	54.3	34.8
			(0.2)	(0.4)	(0.4)	(0.3)	(0.4)
M016801	1990	TSA	16.0	17.5	14.4	19.2	9.6
			(2.1)	(3.1)	(3.0)	(2.9)	(2.9)
	1992	State Assess	16.3	17.4	15.2	19.1	10.8
			(0.2)	(0.3)	(0.3)	(0.2)	(0.3)
M016901	1990	TSA	33.2	34.6	31.6	39.4	21.3
			(2.8)	(3.9)	(3.9)	(3.6)	(4.1)
	1992	State Assess	31.2	33.5	29.0	35.4	23.1
			(0.2)	(0.3)	(0.3)	(0.3)	(0.4)

Table 9

Means, Standard Deviations, Standard Errors of the Means, and Correlations Between NAEP Item Subtests and the Georgia Mathematics Test Sections by Test Form

Statistic	Administered with 1992 Georgia	
	Sum of Items 1 through 9 of Block 7	
	Form 1[a]	Form 4[b]
Mean	5.195	5.280
Standard deviation	2.294	2.265
Standard error	0.016	0.016
Correlation with Georgia Math Test	0.681	0.644
	Sum of Items 10 through 17 of Block 7	
	Form 2[c]	Form 3[d]
Mean	2.234	2.485
Standard deviation	1.658	1.734
Standard error	0.012	0.012
Correlation with Georgia Math Test	0.540	0.538

[a] Form 1 test content order: (1) Social Studies, (2) Science, (3) Health and Safety, (4) Mathematics, (5) Reading, (6) NAEP Items 1-9.

[b] Form 4 test content order: (1) Science, (2) Mathematics, (3) Health and Safety, (4) Reading, (5) Social Studies, (6) NAEP Items 1-9.

[c] Form 2 test content order: (1) Reading, (2) Social Studies, (3) Health and Safety, (4) Science, (5) Mathematics, (6) NAEP Items 10-17.

[d] Form 3 test content order: (1) Mathematics, (2) Reading, (3) Health and Safety, (4) Social Studies, (5) Science, (6) NAEP Items 10-17.

Figure 1

Relative Frequency Distributions of Number Correct Scores on the First Nine Block Seven NAEP Items for the 1990 TSA Results and the 1992 Adminstration of Items Embedded in the Statewide Assessment

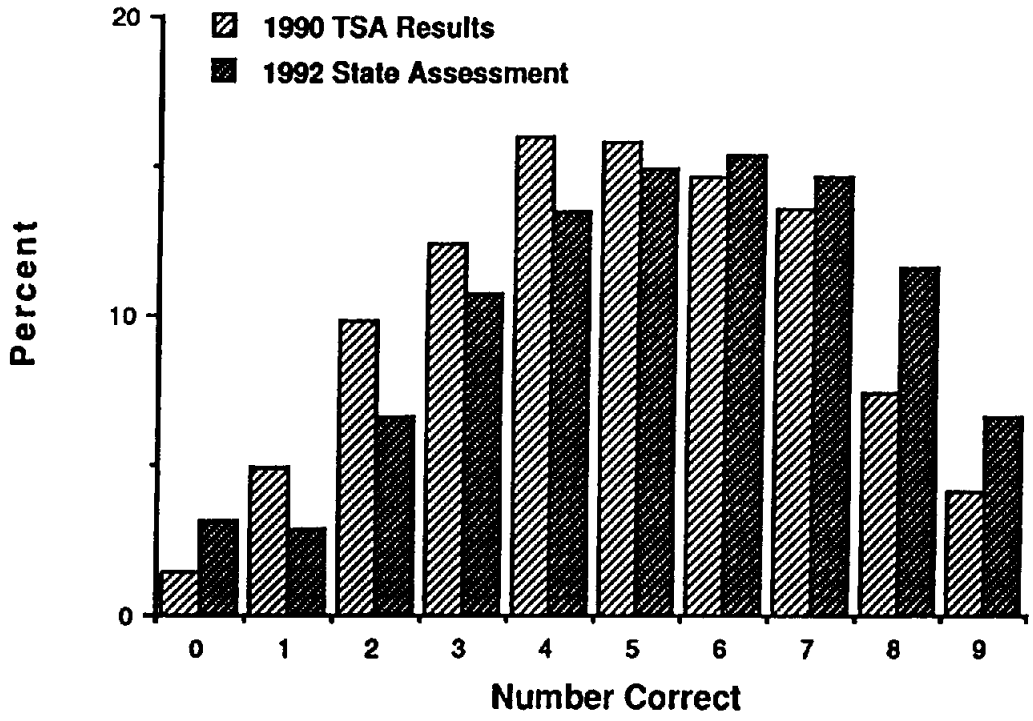


Figure 2

Relative Frequency Distributions of Number Correct Scores on the Last Eight Block Seven NAEP Items for the 1990 TSA Results and the 1992 Adminstration of Items Embedded in the Statewide Assessment

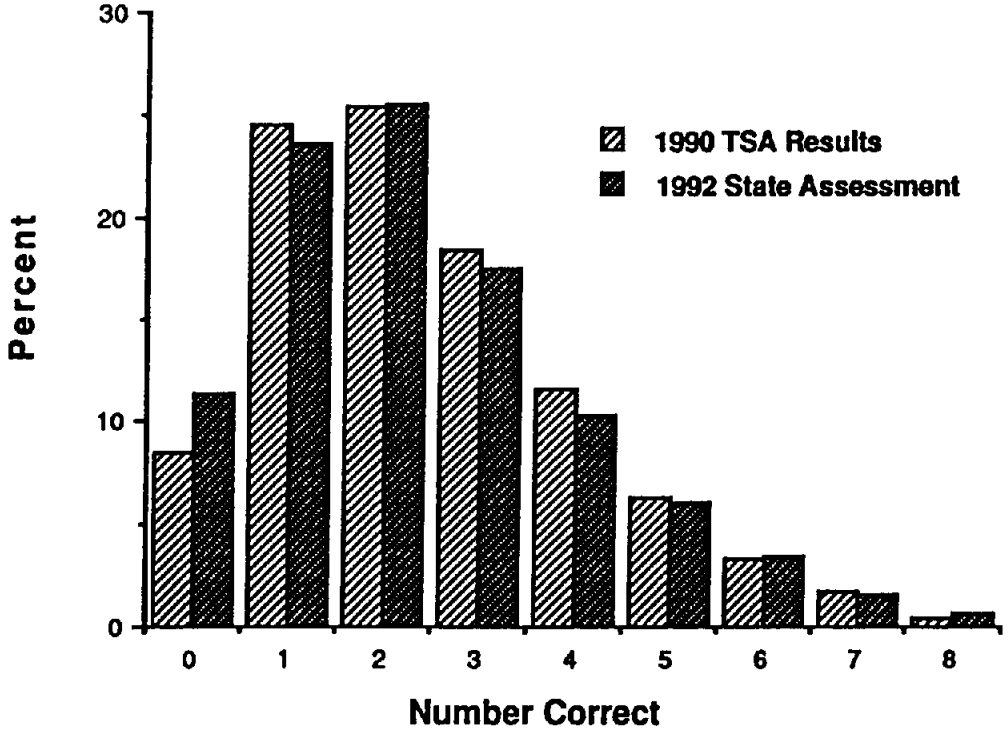


Figure 3

Scatterplot of State Average Percent Correct Values for the 1992 Administration of NAEP Items Embedded in the Statewide Assessment with the State Average Percent Correct Values from the 1990 Administration as Part of the Trial State Assessment

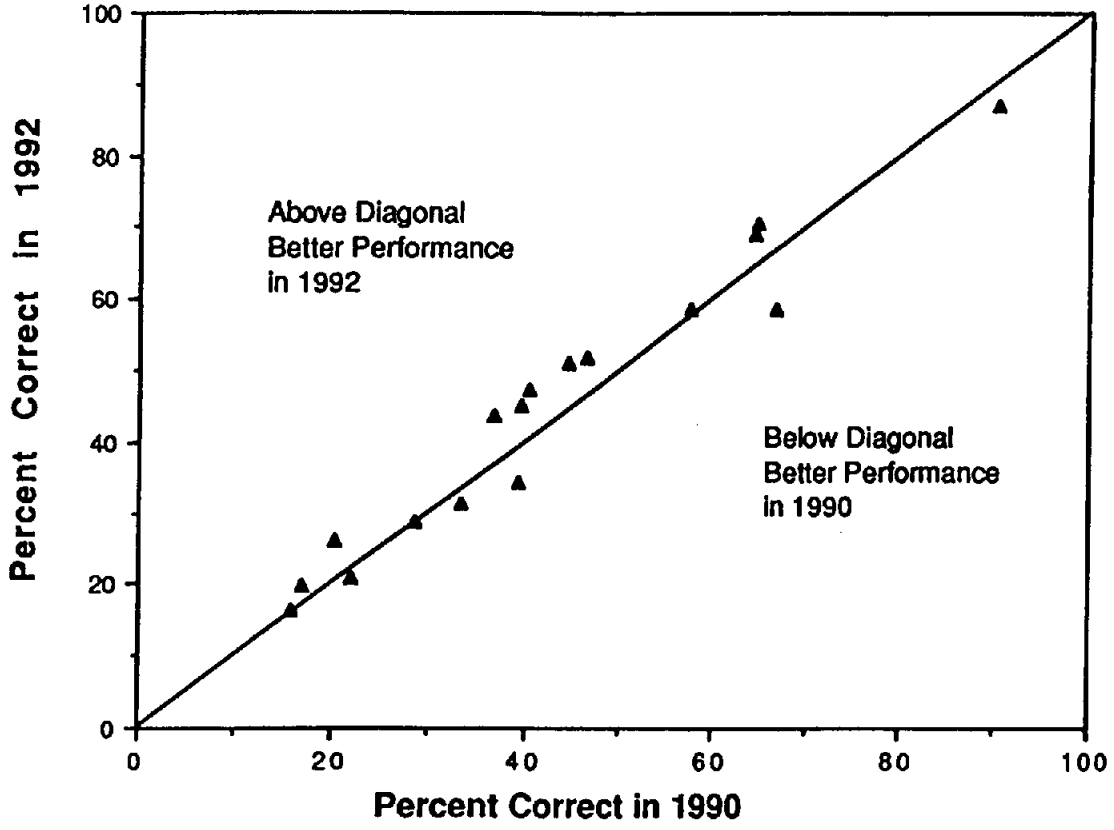


Figure 4

Scatterplot of Change in State Average Percent Correct Values from 1990 to the 1992 Administration of NAEP Items Embedded in the Statewide Assessment with the State Average Percent Correct Values from the 1990 Administration as Part of the Trial State Assessment

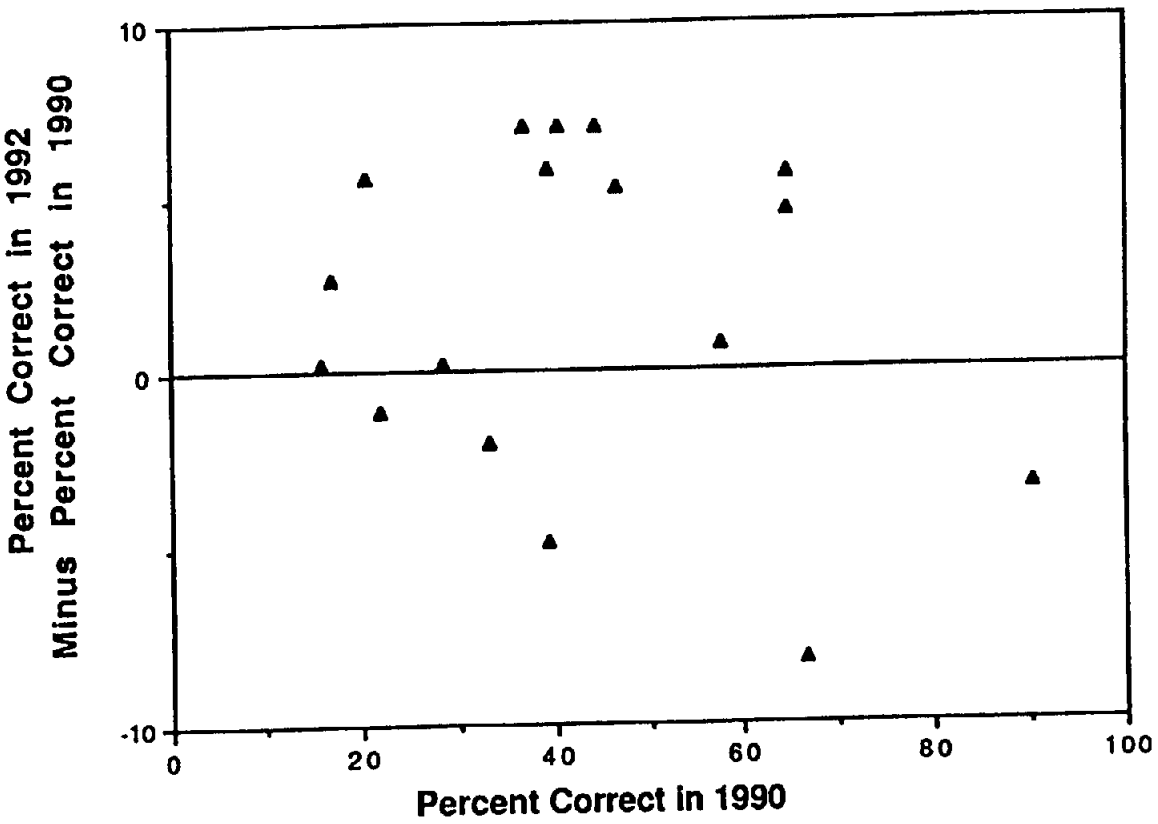


Figure 5

Percent Correct by Year of Administration and Subpopulation For Numbers and Operations Item Number 1 (NAEP Item Label M015501)

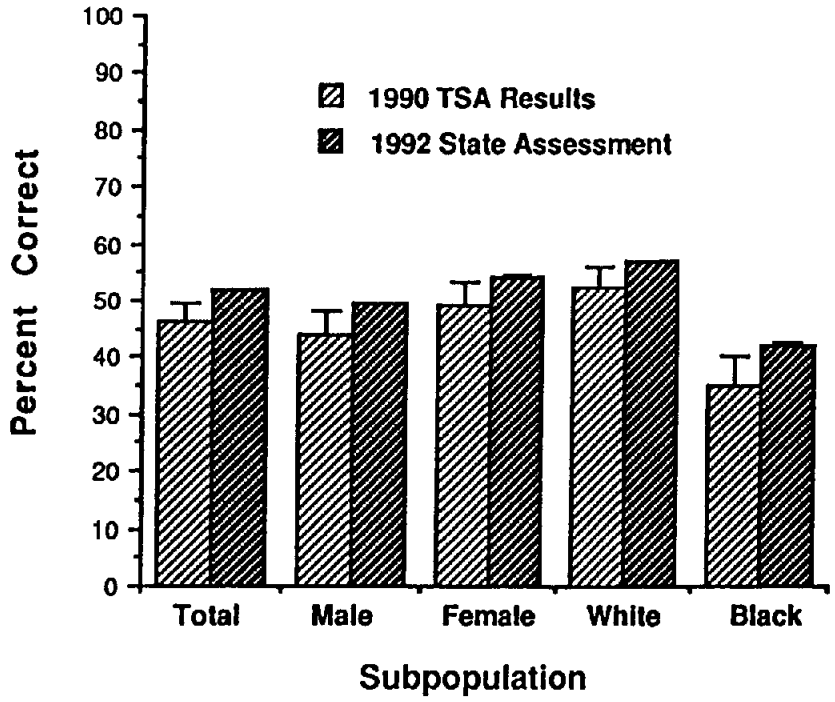


Figure 6

Percent Correct by Year of Administration and Subpopulation For Numbers and Operations Item Number 2 (NAEP Item Label M015901)

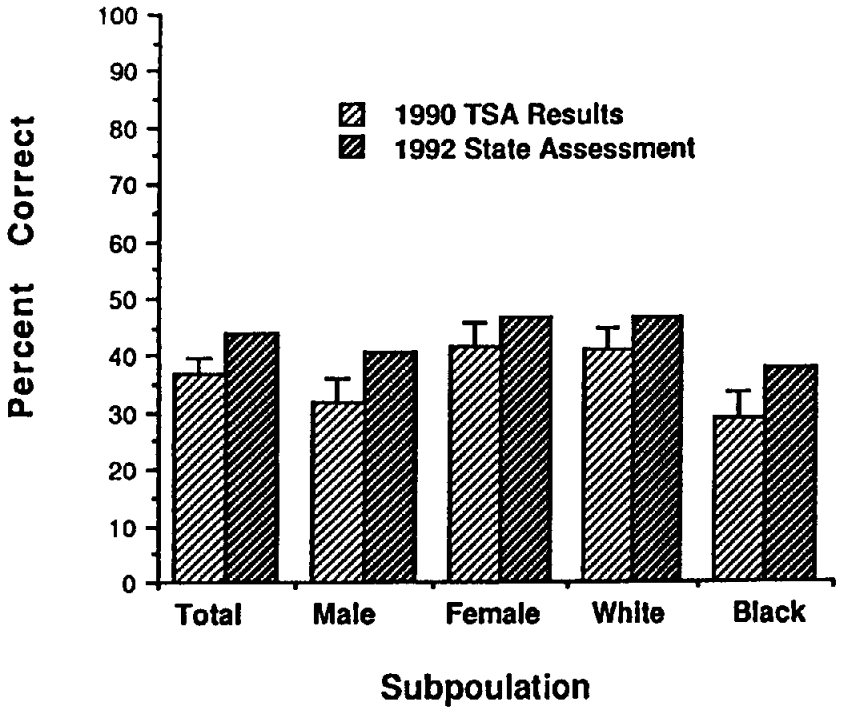


Figure 7

Percent Correct by Year of Administration and Subpopulation For Numbers and Operations Item Number 3 (NAEP Item Label M016501)

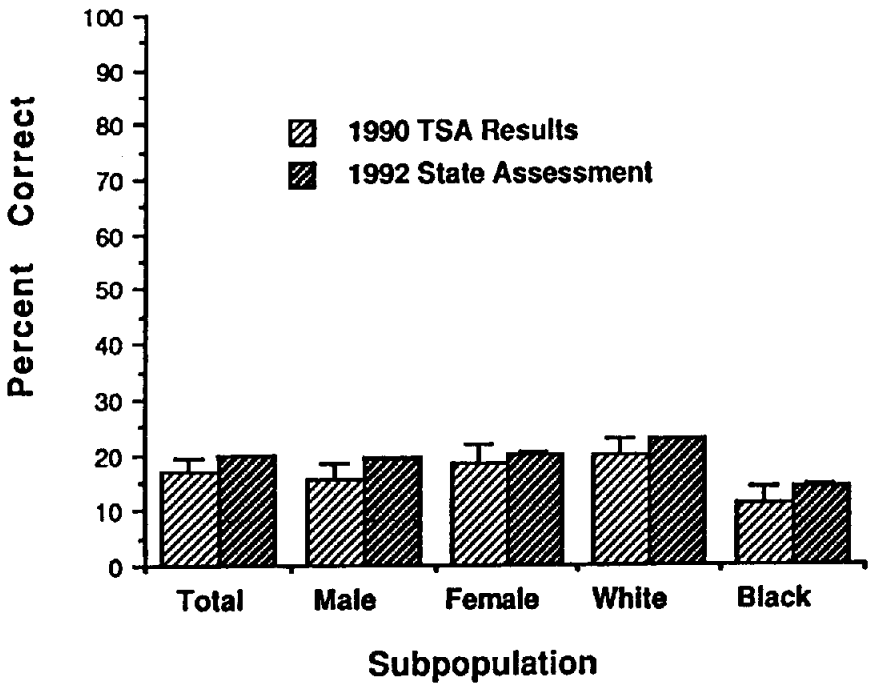


Figure 8

Percent Correct by Year of Administration and Subpopulation For Measurement Item Number 1 (NAEP Item Label M015401)

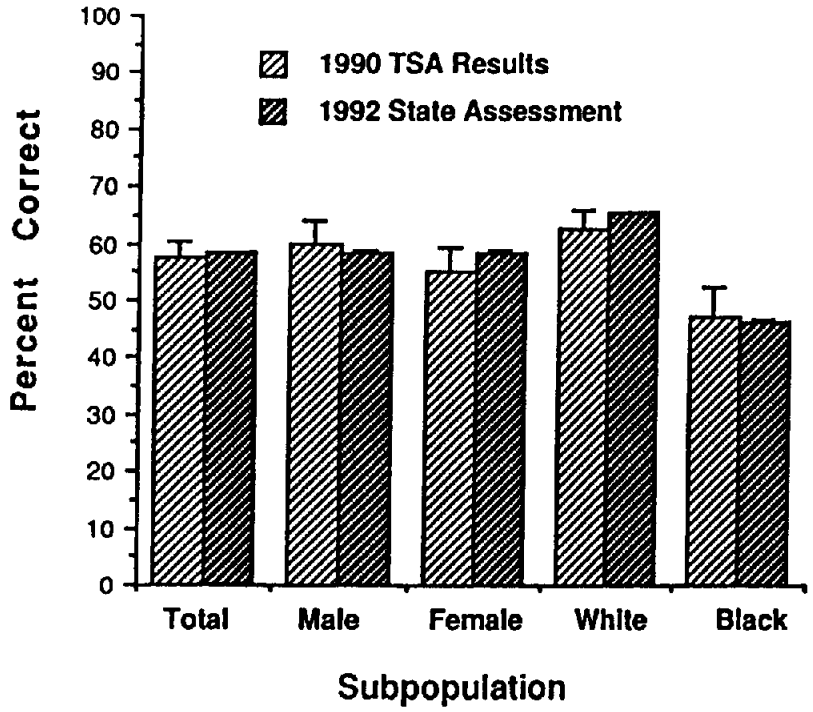


Figure 9

Percent Correct by Year of Administration and Subpopulation For Measurement Item Number 2 (NAEP Item Label M015701)

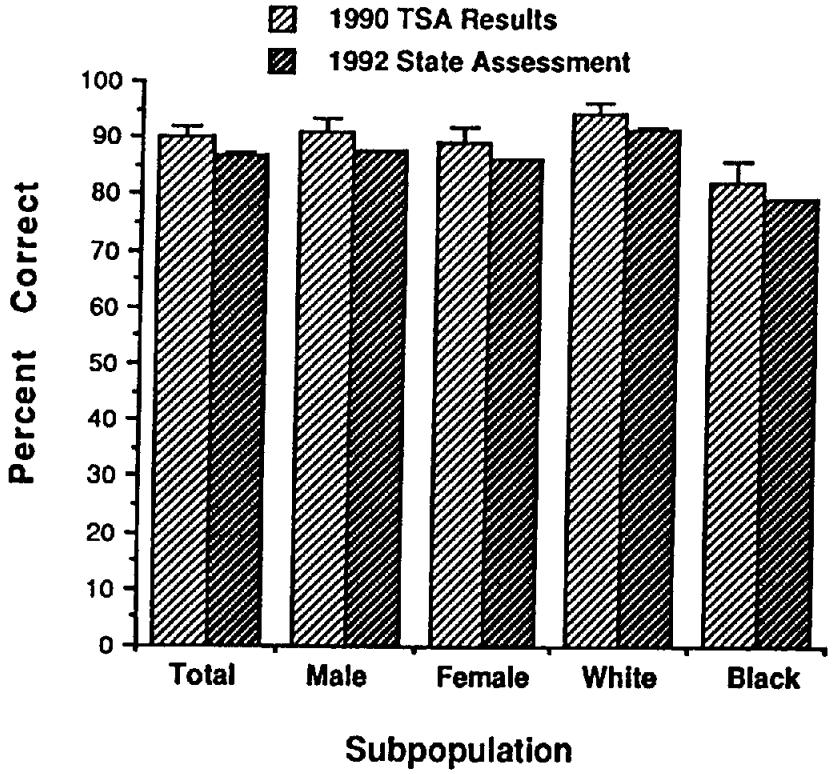


Figure 10

Percent Correct by Year of Administration and Subpopulation For Measurement Item Number 3 (NAEP Item Label M016201)

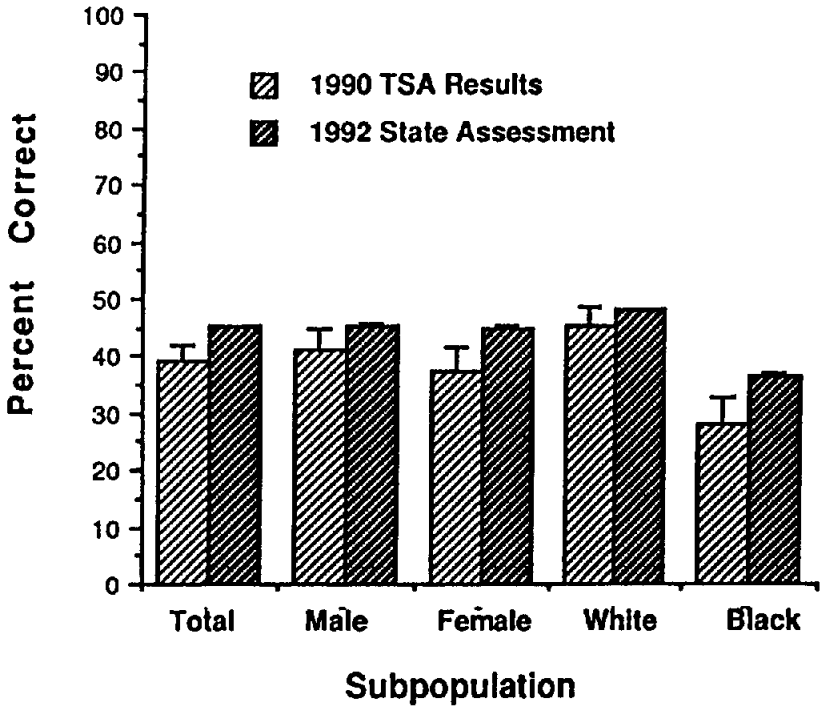


Figure 11

Percent Correct by Year of Administration and Subpopulation For Geometry Item Number 1 (NAEP Item Label M015601)

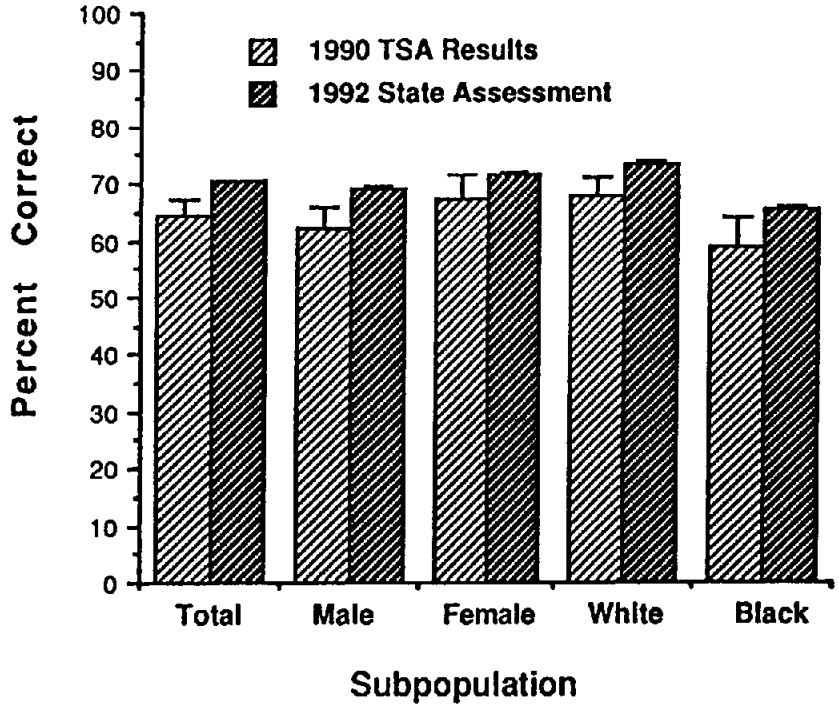


Figure 12

Percent Correct by Year of Administration and Subpopulation For Geometry Item Number 2 (NAEP Item Label M016301)

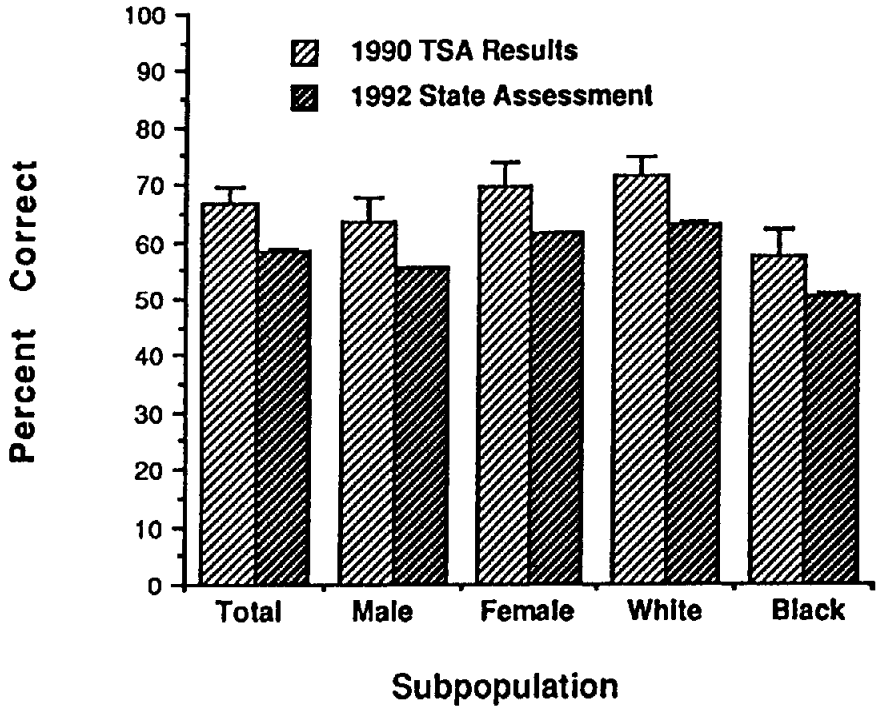


Figure 13

Percent Correct by Year of Administration and Subpopulation For Geometry Item Number 3 (NAEP Item Label M016401)

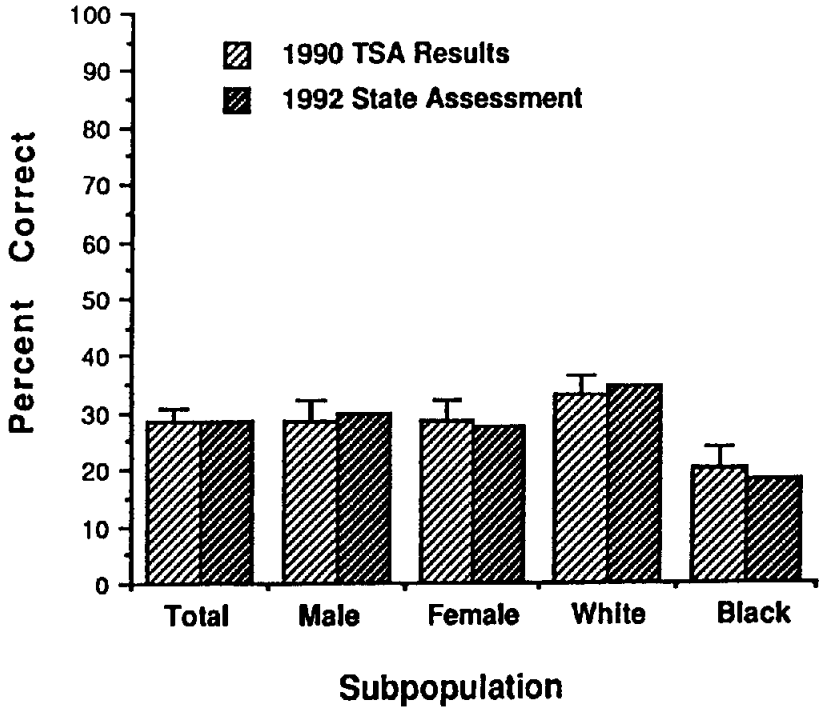


Figure 14

Percent Correct by Year of Administration and Subpopulation For Geometry Item Number 4 (NAEP Item Label M016601)

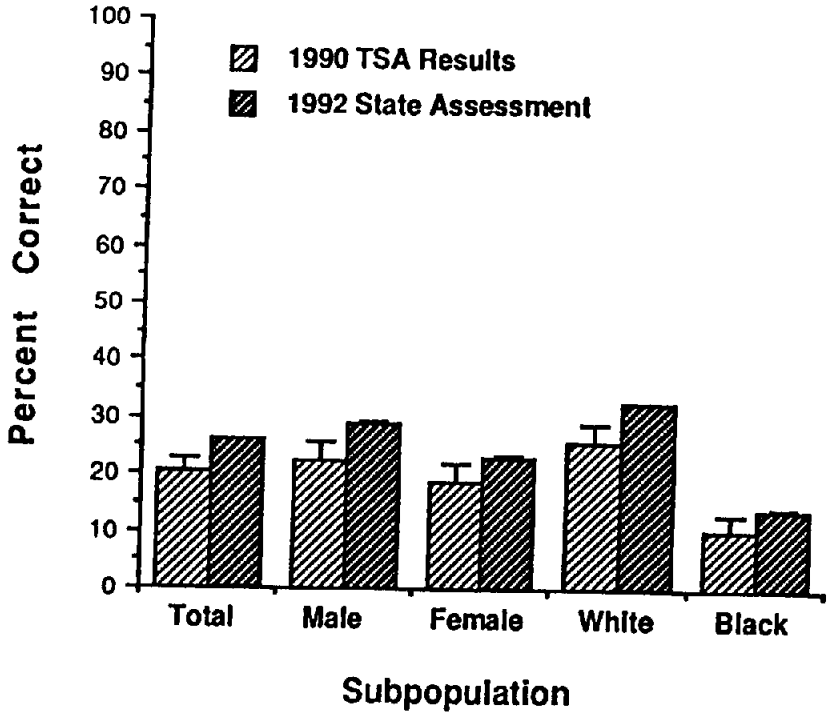


Figure 15

Percent Correct by Year of Administration and Subpopulation For Geometry Item Number 5 (NAEP Item Label M016701)

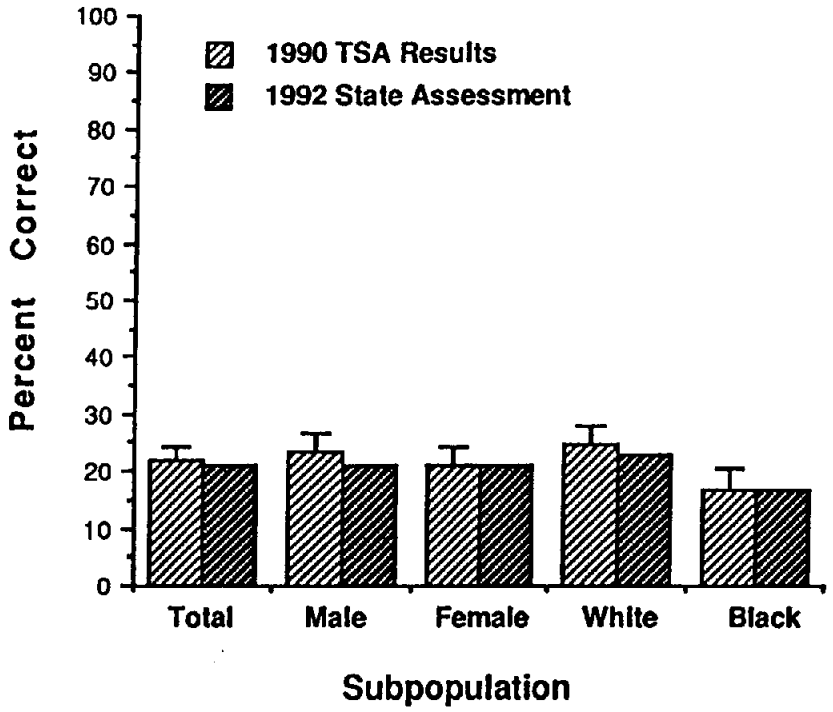


Figure 16

Percent Correct by Year of Administration and Subpopulation For Data Analysis, Statistics and Probability Item Number 1 (NAEP Item Label M015801)

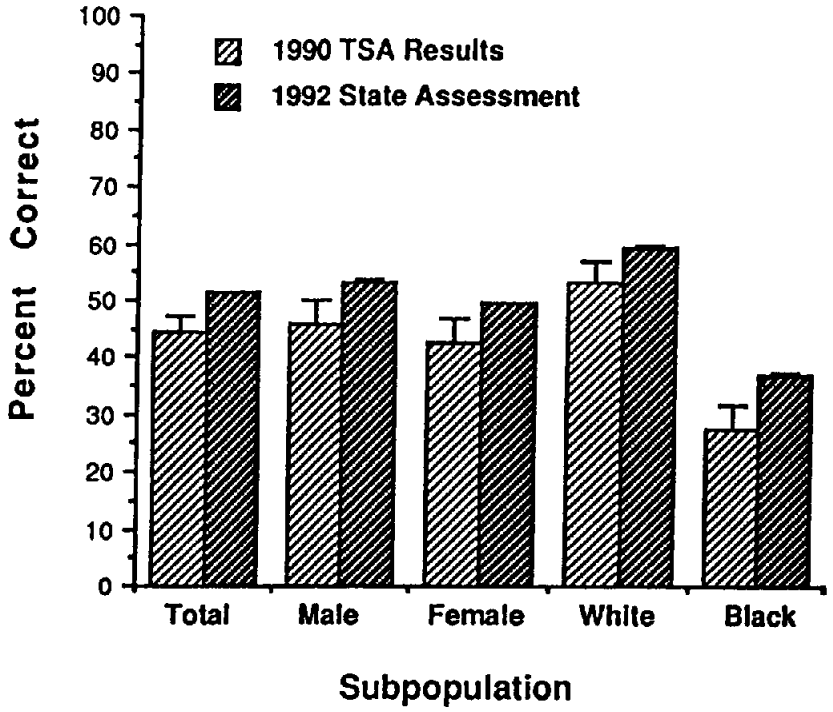


Figure 17

Percent Correct by Year of Administration and Subpopulation For Data Analysis, Statistics and Probability Item Number 2 (NAEP Item Label M016101)

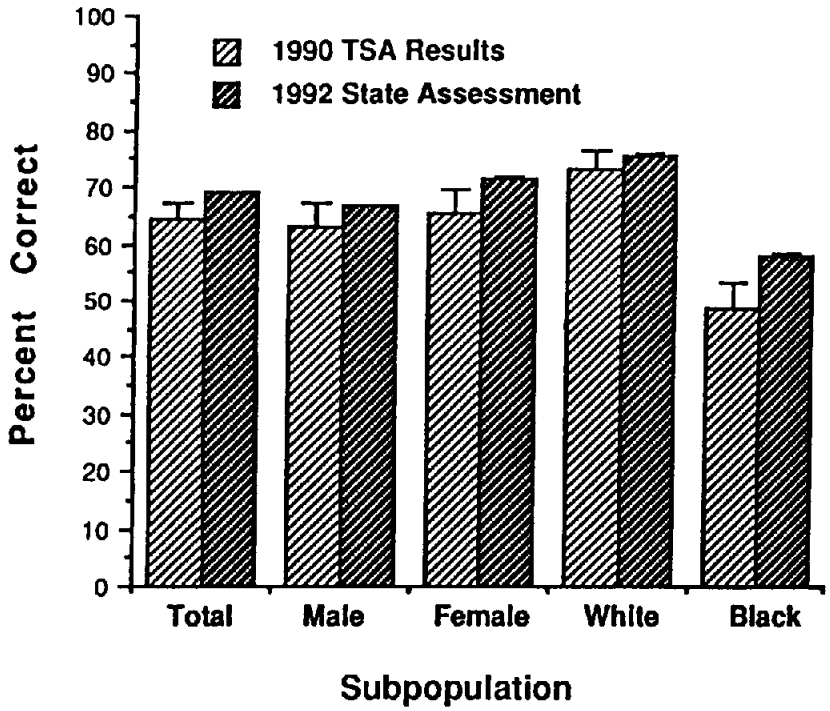


Figure 18

Percent Correct by Year of Administration and Subpopulation For Data Analysis, Statistics and Probability Item Number 3 (NAEP Item Label M017001)

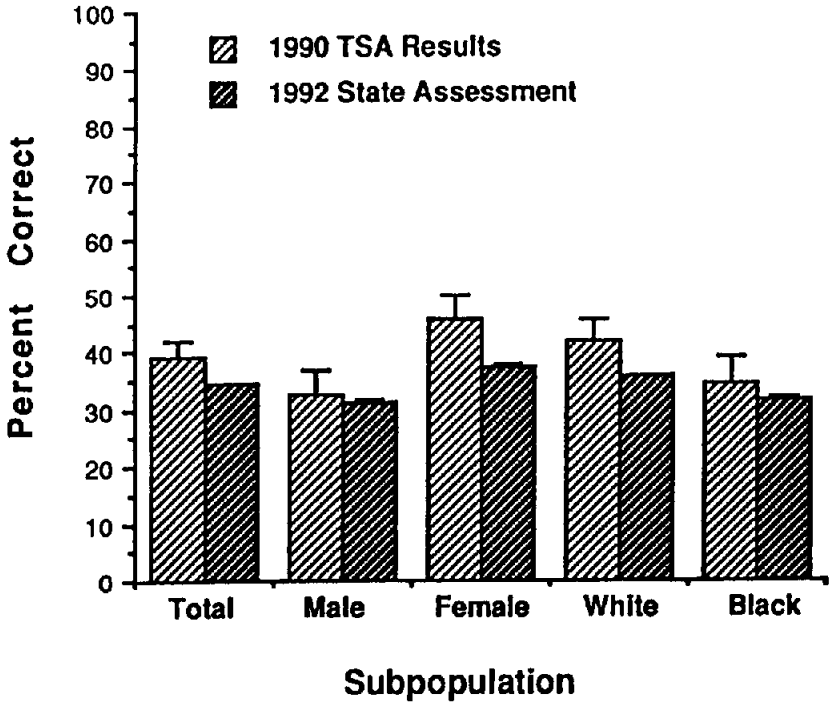


Figure 19

Percent Correct by Year of Administration and Subpopulation For Algebra and Functions Item Number 1 (NAEP Item Label M016001)

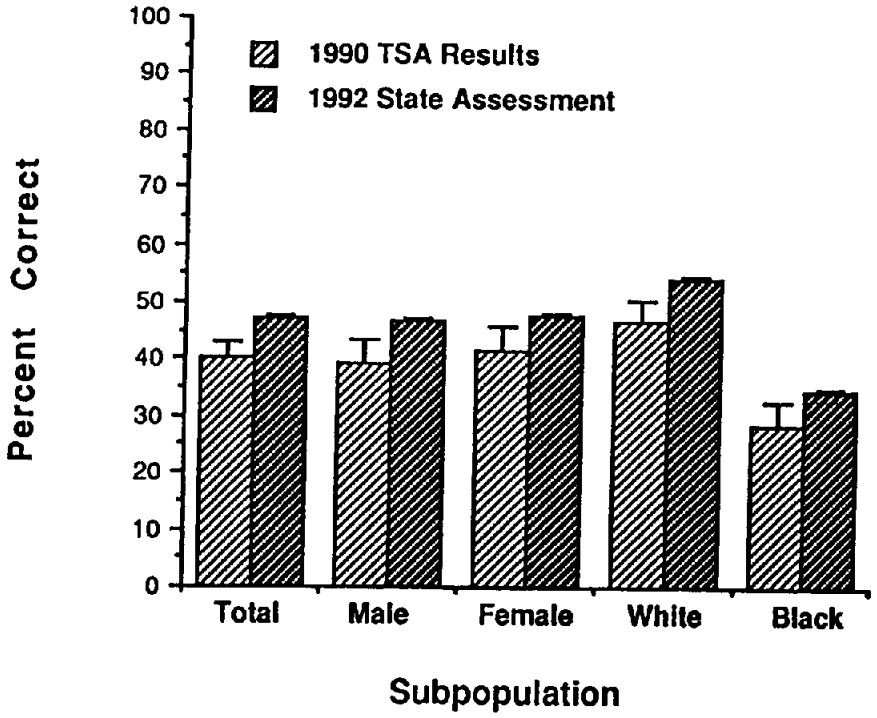


Figure 20

Percent Correct by Year of Administration and Subpopulation For Algebra and Functions Item Number 2 (NAEP Item Label M016801)

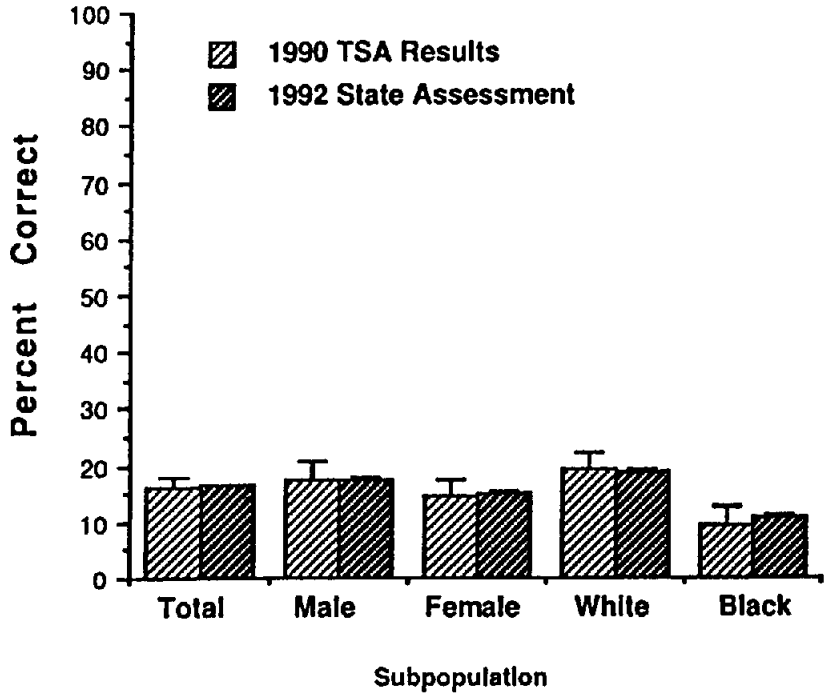


Figure 21

Percent Correct by Year of Administration and Subpopulation For Algebra and Functions Item Number 3 (NAEP Item Label M016901)