



CLOSUP Working Paper Series
Number 17

February 2009

**TEST-BASED ACCOUNTABILITY AND STUDENT
ACHIEVEMENT: AN INVESTIGATION OF
DIFFERENTIAL PERFORMANCE ON NAEP AND
STATE ASSESSMENTS**

Brian A. Jacob
University of Michigan

This paper is available online at <http://closup.umich.edu>

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the Center for Local, State, and Urban Policy or any sponsoring agency

Center for Local, State, and Urban Policy
Gerald R. Ford School of Public Policy
University of Michigan

TEST-BASED ACCOUNTABILITY AND STUDENT ACHIEVEMENT: AN INVESTIGATION OF DIFFERENTIAL PERFORMANCE ON NAEP AND STATE ASSESSMENTS

Brian A. Jacob
University of Michigan

Abstract:

This paper explores the phenomenon referred to as test score inflation, which occurs when achievement gains on "high-stakes" exams outpace improvements on "low-stakes" tests. The first part of the paper documents the extent to which student performance trends on state assessments differ from those on the National Assessment of Educational Progress (NAEP). I find evidence of considerable test score inflation in several different states, including those with quite different state testing systems. The second part of the paper is a case study of Texas that uses detailed item-level data from the Texas Assessment of Academic Skills (TAAS) and the NAEP to explore why performance trends differed across these exams during the 1990s. I find that the differential improvement on the TAAS cannot be explained by several important differences across the exams (e.g., the NAEP includes open-response items, many NAEP multiple-choice items require/permit the use of calculators, rulers, protractors or other manipulative). I find that skill and format differences across exams explain the disproportionate improvement in the TAAS for fourth graders, although these differences cannot explain the time trends for eighth graders.

I would like to thank to Elizabeth Kent and J.D. LaRock for excellent project management and research assistance, and Daniel Koretz for many helpful suggestions. Funding for this project was generously provided by the U.S. Department of Education NAEP Secondary Analysis Grant (#R902B030024). The views expressed in this paper are those of the author and of course, all errors are my own.

© 2007 by Brian A. Jacob. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

1. Introduction

The passage of *No Child Left Behind* (NCLB) in 2001 ensures that test-based accountability will dominate the educational landscape in the near future, making it particularly important for policymakers to understand how and why accountability influences student achievement. The existing evidence is mixed. Several studies suggest that high-stakes testing may be effective at raising student achievement. At the same time, studies of Texas, Kentucky and Chicago have shown that achievement gains on local “high-stakes” assessments dramatically outpaced gains on “low-stakes” exams. While this phenomenon (often referred to as test score inflation) is not new, it has received increasing attention in recent years as test-based accountability programs become more prevalent.

This paper examines the issue of test score inflation in greater depth. The first part of the paper explores the extent to which one particular form of test score inflation actually exists. Specifically, I document the extent to which student performance trends on state assessments differ from those on the National Assessment of Educational Progress (NAEP). While such divergence has been documented in several studies (Klein et al. 2000, Greene et al. 2003, Koretz et al. 1991, Koretz and Barron 1998, Linn 2000), there has been no systematic analysis of this issue nationwide, largely because of the difficulty in collecting state assessment data over a long enough time period and in a format that will allow valid comparisons with the NAEP.¹ For example, many states only report the percent of students in the state reaching a particular benchmark (e.g., percent proficient), which does not allow one to easily trace the trend in average state achievement over time since changes in such categorical ratings can be driven by relatively small changes in underlying achievement.

¹ However, Greene et al. 2003 and Linn 2000 examine more than one school system.

To fill this gap in the research literature, over a period of several years I collected comparable panel of student achievement data on NAEP and state assessments from the limited number of states that administered both state assessments and participated in the state NAEP assessments during the 1990s.). I find evidence of considerable test score inflation in several different states, including those with quite different state testing systems.

In the second part of this paper, I attempt to go beyond previous studies that merely document aggregate performance difference across exams to explore *why* performance trends may differ across these exams. Since state and federal agencies will be using NAEP to benchmark achievement gains on state assessments, it is critical to understand the reasons for differential trends. There are several possible explanations, including differences in student effort across low- vs. high-stakes tests and manipulation of the test-taking pool. The most common explanation, however, involves the differences in the skills and knowledge covered on different exams. All exams differ to some extent in the emphasis they place on various skills and topics. If educators in a state shift their instruction toward the material emphasized on the state assessment, student performance on this measure may increase more rapidly than scores on the NAEP. Hence, in this paper, I examine the extent to which differences in the skills and topics covered across exams can explain the differential performance trends.

To do this, I carefully examine achievement trends in Texas during the 1990s. Between 1996 and 2000, for example, student math performance increased by 0.5 to 0.6 standard deviations on the TAAS compared with roughly 0.1 standard deviations on the NAEP. I utilize detailed, item-level achievement data for the NAEP and the Texas Assessment of Academic Skills (TAAS), which allows me to conduct a number of group test items into categories measuring specific skills and/or knowledge in particular areas and compare achievement trends

within these categories on both the NAEP and the TAAS. This allows me to adjust for content and format differences across exams and estimate how important these factors are in explaining student performance trends.

Several interesting findings emerge from this analysis. First, I find that the differential TAAS improvement cannot be explained by changes in the demographic composition of the test-takers. Second, I find that the differential improvement on the TAAS cannot be explained by several important differences across the exams, including (a) the fact that the NAEP includes open-response items whereas the TAAS only includes multiple-choice items, (b) the fact that many of the multiple-choice items on the NAEP require and/or permit the use of calculators, rulers, protractors or other manipulative such blocks or fraction bars while none of the items on the TAAS do so, or (c) the fact that the TAAS is an un-timed exam and the NAEP is a timed exam. Finally, I find that skill and format differences across exams might explain the disproportionate improvement in the TAAS for fourth graders, but cannot explain the trends among eighth graders.

The remainder of the paper proceeds as follows. Section 2 provides some background and a review of the relevant literature. Section 3 presents the results from the comparison of NAEP and state assessment trends from a variety of states. Section 4 explores the reasons underlying the TAAS-NAEP divergence during the 1990s, and Section 5 concludes.

2. Background

In the past decade, education reform efforts have focused increasingly on standards and accountability-based strategies. To date, nearly all fifty states have succeeded in developing state curriculum standards and have begun to tie student assessments to those standards.

Coupled with these developments has been an overwhelming move toward using student test scores to hold districts schools, teachers and students accountable for learning. Statutes in 25 states explicitly link student promotion to performance on state or district assessments; 18 states reward teachers and administrators on the basis of exemplary student performance, and 20 states sanction school staff on the basis of poor student performance. Many states and districts have passed legislation allowing the takeover or closure of schools that do not show improvement.

The passage of *No Child Left Behind* (NCLB) in 2001 ensures that school accountability will dominate the educational landscape in this country for the foreseeable future. NCLB requires states to test all children in grades three to eight in reading and mathematics in each year, and to report the percentage of students meeting state-defined proficiency levels in each school, breaking down the results by poverty, race, ethnicity, disability and limited English proficiency. Schools will be required to raise the proportion of students meeting these targets each year according to a schedule that ensures that all students are proficient by 2014. If schools fail to make “adequate yearly progress,” they may be subject to increasingly severe interventions, culminating with the closure or reconstitution of the school.

2.1 Conceptual Framework

Incentive theory suggests that test-based accountability will increase student achievement by motivating students and teachers to work harder, causing parents to become more involved in their children’s education and forcing school administrators to implement more effective instruction.² Like other high-powered incentive schemes, however, educational accountability is likely to distort the behavior of individuals (see, for instance, Holmstrom and Milgrom 1991,

² Of course, this view rests on the assumption that these actors have the capacity to respond to the incentives provided by high-stakes testing, which has been questioned by some (Elmore 2002).

Baker 1992, Glaeser and Shleifer 2001). Test-based accountability policies, for example, might lead teachers and administrators to manipulate testing conditions, change the pool of students taking the exam, shift classroom instruction toward tested material, or cheat outright on the exam.

One of the most common critiques of high-stakes testing is that the improvement in student performance is more apparent than real – that rising test scores do not reflect actual increases in knowledge or skill. Consistent with this claim, several recent studies have found that changes in NAEP scores do not match the improvement on local, “high-stakes” assessments in states that have adopted tough accountability policies (Koretz and Barron 1998, Klein et al. 2000). This discrepancy between alternative measures of student achievement is often referred to as test score inflation.

To understand what people mean when they claim that test scores are “inflated” or achievement gains are not “real,” one must first understand something about educational testing. Achievement tests are samples of questions from a larger domain of knowledge. They are meant to measure a latent construct, such as knowledge of mathematics or the ability to read and comprehend written material. The important point is that the score on the test itself is not as important as the inference that can be drawn from the score (i.e., what the test score tells us about the student’s actual set of knowledge and skills). In most cases, we think of the score and the inference as identical. If a student scores high on an exam, he or she must know a lot of math, reading, geography, etc. However, it is easy to think of situations where this might not be true. In the case of cheating, for example, a high score does not necessarily reflect true understanding. When one hears that high-stakes accountability leads to inflated test scores, it means that the test scores are no longer a good indicator of the overall student skills and

knowledge and, by extension, the achievement gains are misleading because they may not reflect a more general mastery of the subject.

There are a several explanations for test score inflation. One is cheating on the high-stakes exam. While such allegations may seem far-fetched, documented cases of cheating have recently been uncovered in a number of states (May 1999, Marcus 2000, Loughran and Comiskey 1999, Kolker 1999) and there is evidence that such test manipulation responds to incentives such as accountability (Jacob and Levitt 2003). A second explanation involves manipulation of the test-taking pool. If teachers or administrators are more likely to exclude low-achieving students from high-stakes exams, then performance on state assessments may appear to increase faster than NAEP. Third, the testing conditions may vary across exams. If administrators ensure better conditions for the state assessment, performance may be higher on this exam. Fourth, student effort may differ considerably across exams. Insofar as the exam has more serious consequences for students and/or schools, it seems likely that students will work harder on state assessments. Finally, performance may differ due to differences in the content of the exams. All exams differ to some extent in the emphasis they place on various skills and topics. If educators in a state shift their instruction toward the material emphasized on the state assessment, student performance on this measure may increase more rapidly than scores on the NAEP. The greater the difference between the exams, the greater difference in performance we would expect to find. Such “teaching to the test” is perhaps the most common explanation for the NAEP-State gap.³

Since state and federal agencies will be using NAEP to benchmark achievement gains on state assessments, it is critical to understand *why* performance trends may differ across these

³ Note that teaching to the test in this context is not the same as what is often referred to as test preparation, which might include activities such as teaching students how to fill properly fill in answer keys or to eliminate unlikely item choices.

exams. The explanations outlined above have starkly different implications for how we assess the viability of test-based accountability. If the gains on local assessments were driven by cheating, manipulation of the test-taking pool, or changes in testing conditions, we would probably not consider them meaningful. In this case, the divergence of performance would be extremely troubling. If the gains were driven by student effort on the day of the exam, we might be more optimistic although this would depend, of course, on whether we think that achievement on the low-stakes exam would have improved given greater student effort.

The case in which the discrepancy is driven by “teaching to the test” is the most difficult to evaluate. It is important to keep in mind that even if an accountability program produced true, meaningful gains, we would not expect gains on one test to be completely reflected in data from other tests because of the inherent differences across exams. Even the most comprehensive achievement exam can only cover a fraction of the possible skills and topics within a particular domain. For this reason, different exams often lead to different inferences about student mastery, regardless of whether any type of accountability policy is in place. For example, simply changing the relative weight of algebra versus geometry items on the NAEP influences the black-white achievement gap (Koretz 2002).

Moreover, there is a range of activities that might be classified as “teaching to the test” which we would evaluate quite differently. On one end of the spectrum, instruction that taught students to answer questions in a very narrow way, focusing perhaps on a particular question format, is less likely to produce meaningful improvement. A classic example of this situation comes from a study of New Jersey state assessment in the 1970’s. Shepard (1988) found that when students were asked to add decimals in a vertical format, the state passing rate was 86 percent, but when they were asked to perform calculations of the same difficulty in a horizontal

format, the passing rate fell to 46 percent.⁴ On the other end of the spectrum, instruction that focuses on a particular set of skills or topics may produce meaningful gains in these areas, but not generalize well to a broader exam such as NAEP. For example, if teachers respond to a low-level state assessment by focusing on basic arithmetic, student performance may increase more rapidly on the this exam than on the NAEP. Yet, we might still believe that the students have made important progress in the area of arithmetic.

Hence, knowing the reason(s) student performance trends differ across exams will not always tell us exactly what “should” be done. It will, however, provide some insight into learning under accountability and thus inform policymakers in making decisions about various aspects of education reform.

2.2 Literature Review

A growing body of evidence suggests that test-based accountability programs increase student achievement (Richards and Sheu 1992, Grissmer et. al. 2000, Deere and Strayer 2001, Jacob 2002, Carnoy and Loeb 2002, Hanushek and Raymond 2002).⁵ Specifically, researchers have found that student achievement in math and reading increased substantially following the introduction of high-stakes testing policies in South Carolina (Richards and Sheu 1992), Texas (Deere and Strayer 2001) and Chicago (Jacob 2002). Several national studies utilizing state

⁴ The author suggests that this differential was due to the fact that schools had traditionally taught students decimal arithmetic in a vertical format, and they really did not understand the concept well enough to translate this to a different format. If this were the case, one might conclude that the achievement represented by the vertical scores was not meaningful. This example, however, also demonstrates the difficulty inherent in such judgments. One might argue that the format differences represented by vertical and horizontal arithmetic entail real differences in required skills and difficulty. Horizontal problems require a student to understand place value well enough to know to center on the decimal point rather than right-justifying the decimals when rewriting the problems vertically (in their mind or on paper).

⁵ There is also evidence that such accountability programs, particularly mandatory high school graduation exams, increase dropout rates among low-achieving students. See, for example, Jacob (2001), Lillard and DeCicca (2001) and Dee (2002). While these findings are obviously critical in evaluating the overall effect of accountability, I do not discuss these studies here since they are not central to the research question in this project.

NAEP data have found similar results (Grissmer et. al. 2000, Loeb and Carnoy 2002, Hanushek and Raymond 2002).⁶ For example, even after controlling for changes in exclusion rates, Loeb and Carnoy (2002) found a significant, positive relationship between the strength of a state's accountability system and 8th grade math achievement gains from 1996 to 2000, with particularly large effects for black students.⁷

At the same time, several studies have documented that student performance on NAEP may differ substantially from achievement on high-stakes, state assessments, suggesting that the achievement gains observed under accountability may not be generalizable. For example, Koretz and Barron (1998) found that students in Kentucky improved more rapidly on the state's exam, KIRIS, than on other assessments. Between 1992 and 1994, for example, KIRIS scores in fourth-grade mathematics increased by about 0.6 standard deviations in contrast to NAEP scores, which increased 0.17 standard deviations. Moreover, the NAEP gains were roughly comparable to the national increase and not statistically different from gains in many other states. Klein et. al. (2000) conducted a similar analysis of Texas, comparing performance trends on the Texas state assessment (TAAS) and the NAEP. The researchers in this case found that TAAS scores increased by considerably more than NAEP scores. For example, in 4th grade reading, TAAS scores for black students increased by roughly 0.50 standard deviations while NAEP scores only increased by 0.15 standard deviations.⁸

⁶ The primary advantages of these studies are that they are able to examine a wider sample of jurisdictions and they utilize a common outcome measure that should be less subject to manipulation than local assessments. The disadvantage of such studies is that it is difficult to separate the effect of the accountability from that of other state policies.

⁷ Amrein and Berliner (2002) found little relation between high-stakes graduation exams and NAEP math and reading achievement in 4th or 8th grade. Unfortunately, their analysis is largely descriptive and because the authors present results separately for each state, and consider a variety of different time periods, it is quite difficult to evaluate their validity of their analysis, assess the statistical significance of their findings, or compare their results to those in other studies.

⁸ More generally, researchers have found considerable evidence of test score inflation throughout the country during the past two decades. In 1987, Cannell (1987) discovered what has become known as the "Lake Wobegon" effect—

However, it is not clear whether cases such as Texas and Kentucky are representative of a more general phenomenon. Greene et al. (2003) compare student scores on high- versus low-stakes exams in two states and seven districts. The authors find a high correlation between test score levels across all jurisdictions, but weaker correlations in test score gains. Moreover, there was considerable heterogeneity across jurisdictions. For example, the correlation between annual gains on the high- and low-stakes exam was 0.71 in Florida but only 0.17 in Virginia. The seven districts generally had low correlations between gains. Unfortunately, the analysis suffers from several technical shortcomings. Most importantly, the authors claim that a high correlation between gains on a high-stakes and low-stakes exam implies the absence of test score inflation. The correlation, however, is largely picking up the relative ranking of schools within a district. It is possible for the correlation to be high even if all schools in the district made substantially smaller gains on the low-stakes test than on the high-stakes test.

More importantly, there has been little research on the reasons *why* student performance differs between NAEP and local assessments. There is a considerable body of evidence on strategic responses to test-based accountability suggesting that such policies lead to an increase in test-exclusions (Cullen and Reback 2002, Figlio and Getzler 2002, Jacob 2002), teacher cheating (Jacob and Levitt 2003), other testing conditions (Figlio and Winicki 2002) and test preparation (Tepper et. al. 2002, McNeil and Valenzuela 2001). However, this work does not

the fact that a disproportionate number of states and districts report being “above the national norm.” This phenomenon was documented in several studies (Linn et. al. 1990, Shepard 1990). Similarly, Linn and Dunbar (1990) found that states have made smaller gains on the National Assessment of Educational Progress (NAEP) than their own achievement exams. One of the earliest studies on this topic examined score inflation in two state testing programs where accountability policies were introduced in the 1980s (Koretz et. al. 1991). In this study, researchers administered one of two independent tests to a random selection of elementary classrooms—a commercial multiple-choice test comparable to the high-stakes exam used in the states or an alternative test constructed by the investigators to measure the same content as the high-stakes test. A parallel form of the high-stakes test, designed by the publisher, was also administered to an additional randomly selected group of classes. Results from the actual high-stakes exam and the parallel form were compared to assess the effect of motivation while results from the two independent exams and the actual exam were compared to examine the generalizability of learning. They found considerable evidence of score inflation, particularly in math.

explicitly exam the issue of test score inflation. Jacob (2005) makes an effort to fill this gap. He found that the introduction of an accountability policy in the Chicago Public Schools resulted in large gains on the high-stakes exam – the Iowa Test of Basic Skills (ITBS) – but relatively smaller gains on the state-administered, low-stakes exam, the Illinois Goals Assessment Program or IGAP (Jacob 2005).⁹ In an effort to explore whether this might be due to a shifting and/or narrowing of the curriculum, Jacob examined item-level achievement trends on the ITBS. For the math exam, he found that student achievement gains were substantially larger in areas such as computation and number concepts, topics that comprise a relatively greater portion of the ITBS than the IGAP. In reading, on the other hand, student achievement was equivalent across skill and topic areas. This evidence suggests that shifts in curriculum may have played a large role in the differential math performance, but are unlikely to explain such gaps in reading.

3. A Comparison of NAEP and State Assessment Trends

In this section, I document NAEP-state assessment trends in several different states. I take particular care to construct comparable test data across a period of time to allow the most valid comparisons possible.

3.1 Data

The data for this analysis consists of the statewide average achievement level on the NAEP or state assessment by subject, year and grade. The NAEP data was obtained directly through the Web Tool on the NCES website. The state data was obtained by contacting the state assessment offices in each state.

⁹ An important exception was eighth grade, in which Jacob (2005) found comparable gains on the ITBS and IGAP.

There are several advantages of using the NAEP as the low-stakes comparison. First, during the period of this analysis, the NAEP was truly a low-stakes test – not all students in a state took the exam, individual student-level results were not reported, and the results had no meaningful effect on teachers, administrators or students. Second, it is a very broad measure of achievement in a particular subject, in part due to the matrix sampling design of the test. Third, the test is very consistent from year to year.

There are also advantages of looking at the state level as opposed to district level. The state is the key actor under NCLB, and much education policy is set at the state level. Moreover, examination of state level results should reduce measurement error. There is also a benefit of looking over a long period of time so that any observed changes are less likely to be driven by idiosyncratic fluctuations in cohorts or other factors.

However, there are a number of difficulties in comparing achievement over time across different exams. The first concern involves the comparability of exams over time. While the NAEP is constructed explicitly to allow comparisons over time, many states have changed their assessment system over the past decade. The most common changes include (1) switching to a new version of a nationally normed standardized exam (e.g., from the CTBS4 to the new version of the CTBS called the TerraNova), (2) switching from a norm-referenced test (NRT) to a criterion-reference test (CRT), (3) switching from one NRT to another NRT (e.g., from the ITBS to the SAT-9); (4) switching from one CRT to another CRT. The change to a newer version of the same NRT is less problematic because the test publishers generally conduct equating studies that allow one to create a crosswalk between, for example, the CTBS4 and the TerraNova.¹⁰ Similarly, some states such as North Carolina that introduce new CRTs provide means of comparing scores with older exam results. However, in other cases, it is difficult to draw

¹⁰ Even these equating studies are subject to critique.

rigorous conclusions from comparisons across different exams. For example, Maine introduced a new CRT in 1999 which they did not equate with the state exam administered between 1994 and 1998. It is therefore impossible to compare student achievement trends on state assessments before and after 1999.

A second concern involves the reporting metric. The NAEP provides an average scale score for the state along with a student level standard deviation that allows one to calculate the annual change in student performance in terms of standard deviation units (i.e., effect sizes). Unfortunately, many states do not report results in this manner. Indeed, many states do not even publish state level averages of the underlying raw or scaled score, but rather report student performance in terms of the percent meeting various proficiency levels.¹¹ Among those states which did have measures of the underlying test scores, often we were only able to obtain average national percentiles or normal curve equivalents.

A final concern involves the comparability of grades. The state NAEP is given in the 4th and 8th grade, so the best comparison will be with state exams administered to the same grades. In most states, this is possible. Moreover, in all states the achievement trends on the state exams are roughly comparable across grades.

3.2 Sample

Given the data limitations described above, I was only able to obtain the appropriate data for four states – Texas, North Carolina, Arkansas and Connecticut. In theory, it should be

¹¹ While this metric may be useful for some policy purposes and is certainly a convenient way to report information to the public, it creates difficulties for determining the change in student performance over time, both because the cutoffs for proficiency levels may change over time but also because small changes in underlying performance may lead to larger changes in classification from one category to another. The American Institutes of Research (AIR) has compiled much of this easily available information in a state assessment database. This data is available at www.schooldata.org.

possible to conduct a comparable analysis for approximately 11 states total. Unfortunately, state departments of education in these additional states were not able to provide the type of data that would allow rigorous comparison across NAEP and the state exams. On the other hand, the four states included in this section of the analysis reflect a range of state testing regimes, from Texas which had basic skills exam during the analysis period, to North Carolina, which had an exam that was supposedly constructed to be much more similar to the NAEP.

For each state, I first create normalized state average scores within each subject-grade by subtracting the state average for the first year the exam was administered and dividing by the student-level standard deviation of the exam in that year. Specifically, we

calculate $y_{ejsgt} = \frac{\tilde{y}_{e,j,s,g,t} - \tilde{y}_{e,j,s,g,t=1}}{s_{e,j,s,g,t=1}}$, where \tilde{y} reflects the raw score and y reflects the

standardized score. Each of the states included in the analysis provided state mean scale scores and student-level standard deviations.

3.3 An Initial Look at State Achievement Trends

In order to provide a simple and transparent picture of student achievement trends on NAEP versus state exams, Figures 1 to 4 present graphs of achievement trends in key grades by state and subject. We first consider Texas, which has received much attention for having substantial test score inflation. Figure 1 shows NAEP and TAAS achievement trends for 4th and 8th graders in math and reading. While NAEP math scores appear to have increased steadily over the past 15 years, it is clear that TAAS math scores increased even more rapidly over the time period in which both exams were administered. The pattern is even starker for reading, where NAEP shows almost no gains over the 12 year period.

Figure 2 presents comparable trends for North Carolina. This provides an interesting contrast to Texas because, unlike the TAAS, the North Carolina exam was supposedly much closer to the NAEP in design and content. In math, we see substantial gains on the state exam during this period. The NAEP shows large and steady gains from the early nineties onward, but the pace of improvement is not quite as large as on the state exam. For reading, the patterns in North Carolina mirror those in Texas, although again somewhat muted – substantial gains on the state assessment compared with little, if any, gains on the NAEP. One exception is fourth grade NAEP reading, which shows sizeable gains since 1998.

Figures 3 and 4 show the trends for Connecticut and Arkansas respectively. For both states, the time period in which both state and NAEP exams were administered is limited. The patterns for math achievement in Connecticut are similar to those in North Carolina – the rate of growth of the state exam outpaces the rate of growth of NAEP over the comparable time period. The reading scores on the Connecticut state exam are somewhat noisy, but there is some

evidence that NAEP gains outpaced state exam exams during the mid-1990s in fourth grade. In Arkansas, state assessment and NAEP tell a similar story for reading – namely, remarkably little progress over the nearly 15 year period. In math, the NAEP and state assessment are roughly comparable, although there is a limited time period over which they can be compared.

3.4 Estimation Results

In order to better quantify the results hinted at above, we now present some simple regression estimates using the data above. The unit of analysis is a state*subject*grade*year*exam type, where exam type is either NAEP or the state’s own assessment. We limit our sample to time periods in which both exams were administered.¹² We estimate the following regression separately for math and reading:

$$(1) \quad y_{egst} = \beta_1 StateExam_e + \beta_2 Year_j + \beta_3 (Year_t * StateExam_e) + f(ExamYear_{egst}) + \delta_{gs} + \epsilon_{egst}$$

where y_{egst} reflects the average score on exam e in grade g , state s and year t . The variable $StateExam$ is a binary indicator that takes on the value of one if the observation reflects a state exam score ($e = state$) and zero if the observation reflects a NAEP exam score ($e = NAEP$). The variable $Year$ is a continuous variable that reflects the calendar year. The model includes a cubic in the number of years the particular exam had been administered, denoted $ExamYear$, in order to control for “practice” effects – namely, that students may do poorly on an exam, regardless of type, during early administrations, before teachers and school officials have had the opportunity

¹² Specifically, we include the following observations: Texas, Math – 1996 to 2003; Texas, Reading, Grades 3 to 5 – 1994 to 2003; Texas, Reading, Grades 6 to 8 – 1998 to 2003; North Carolina, Math – 1992 to 2002; North Carolina, Reading, Grades 3 to 5 – 1994 to 2002; North Carolina, Reading, Grades 6 to 8 – 1998 to 2002; Arkansas, Math, 2002 to 2003; Arkansas, Reading – 1998 to 2003; Connecticut, Math – 1996 to 2000 (starting in 2001, CT administered a different version of its state exam that is not comparable to the exam administered between 1994 and 2000); Connecticut, Reading, Grade 3 to 5 – 1994 to 1998; Connecticut, Reading, Grades 6 to 8 – None (because the 8th grade NAEP reading exam was first administered in CT in 1998 and then again in 2002 while the state test data ends in 2000).

to see the exam and adjust their curriculum/pedagogy accordingly. Finally, the model includes a set of fixed effects for state*grade group, δ_{gs} , where grades are placed in one of two groups: grades 3 to 5 or grades 6 to 8.

The coefficient β_2 reflects the average annual learning gain on the NAEP exam and β_3 captures the additional annual increment on the state assessment. The hypothesis of test score inflation suggests that β_3 should be positive. The inclusion of state*grade fixed effects, α_{sg} , insures that our estimates are not driven by compositional changes in the particular grades and/or states included in the analysis at different points in time. In order to account for serial correlation, or other unobserved heterogeneity in the error terms within states, we calculate the standard errors using a block bootstrap (1000 replications) where the blocking variable is state.

Table 1 presents the results of this estimation. Column 1 in the top panel indicates that NAEP math scores increased roughly .048 standard deviation a year in these states over this time period, but that state assessment scores grew almost twice as fast with an annual average gain of approximately .088 standard deviations per year. Columns 2-5 show the results estimated separately for grade and year groupings. While the standard errors are sufficiently large to preclude rigorous comparisons across these separate samples, the point estimates suggest that differential state assessment growth – i.e., test score inflation – was much larger from 1992 to 2000 relative to 2000 to 2003. This is consistent with the increasing state emphasis on the NAEP exam, particularly since the passage of NCLB. Interestingly, it also appears that test score inflation was considerable larger among older elementary students – that is, in grades 6 to 8, relative to grades 3 to 5.

The bottom panel presents the results for reading. Overall, test score inflation appears to be relatively much more pronounced in reading – due primarily to the small improvements in reading performance on the NAEP exam in these states over this period. Unlike in math, the degree of inflation appears comparable across grade levels. And there is simply not enough information to say anything meaningful about differences across time periods.

4. What was driving the NAEP-TAAS Divergence in the 1990s?

The results presented above suggest that there has been considerable test score inflation in several states besides Texas, even states with quite different testing systems such as North Carolina. Having documented some of the aggregate performance differences for several states, this section explores the factors that might be driving differential achievement trends. As

discussed above, there are a variety of different reasons that student performance trends may differ across exams. Here I seek to examine several possible factors, including the composition of the test-takers, the conditions of the exam (specifically whether it is timed or not), and most importantly, differences in question content and format. The NAEP is designed to test a broad range of skills and knowledge, with the framework developed by national organizations and items written by teachers across the country along with subject specialists at the Educational Testing Service (ETS). State assessments, in contrast, differ widely in the skills and knowledge they measure, and the topics they emphasize. If educators in a state shift their instruction toward the material emphasized on the state assessment, it is likely that student performance on this measure will increase more rapidly than scores on the NAEP.

By examining how student performance trends vary across topics, we can gain some insight into the generalizability of achievement gains under the current accountability policies. For example, if we find that differences in test content do not explain much of the differences in student performance across exams, we would conclude that such differences were due to other factors such as student effort, test exclusions or perhaps even cheating. On the other hand, we may find that the rapid gains on state assessments (relative to the NAEP) are driven, for example, by improvements in particular topics such as measurement or data analysis. This information will inform educators and policymakers as they make decisions about the direction of education reforms in their state. For example, state officials may conclude to adjust curriculum guidelines and/or professional development to focus more on topics such as estimation or spatial reasoning.

4.1 Data

Analyzing the role of question content and format on student performance presents unusual data challenges. In order to examine achievement trends on specific skills or topics, it is necessary to explore item-level achievement data. Moreover, in order to assess the content and format of particular items and to identify comparable items across different exams, it is necessary to review the exam items themselves. For this reason, the analysis here is limited to a single state – Texas.

This section makes use of detailed student- and item-level achievement data from the NAEP and the Texas Assessment of Academic Skills (TAAS). The restricted-use NAEP data provides item-level achievement data for individual students along with a host of other information on the students, teachers and schools included in the sample. Because items in the NAEP are reused, access to the actual test items is closely monitored. NCES releases a small number of items periodically, after which time the items are removed from the item pool for subsequent exams. With special permission from NCES, I was able to review the items from previous NAEP exams (including items that are currently in use). To complement the NAEP data, I obtained data from the Texas Education Authority that includes both student level achievement data as well as copies of the actual exams. The data on the TAAS includes how each student answered each item, along with key demographic variables.

My sample includes all students who took the NAEP or TAAS math exam in grades 4 or 8 in the years 1996 or 2000. I limit my analysis to these grades and subjects since these were the only students to whom both the NAEP math exam was administered. I limit the analysis to math

since it is more reasonable to assume that one can categorize math items into skill/topic categories, whereas the same exercise is considerably more difficult for reading comprehension.

In constructing the TAAS sample, I drop all students who are missing math scores (roughly 17 percent in each grade-year), which can be due to either absence on the day of the exam or exemption from testing due to placement in particular bilingual or special education programs. For ease of computation, I then take a random 5 percent sample within each grade-year.¹³ To create the NAEP sample, I start with all public school students who took the NAEP exam in Texas. I then exclude roughly 6 percent of students who were not included for reporting purposes by NAEP.¹⁴

Table 2 presents descriptive statistics on the sample.¹⁵ The first row shows the average percent of multiple-choice items that were answered correctly on each exam in each year. The average student answered roughly 70-80 percent of the multiple-choice items correctly on the TAAS compared with only 50-60 percent on the NAEP. The second panel shows basic demographic characteristics. Roughly 15 percent of elementary students in Texas are Black, 35 percent are Hispanic and 40-50 percent of students are eligible for free- or reduced-price lunch.

4.2 Grouping Items into Skill and/or Knowledge Categories

In order to determine the degree to which differences in performance between the NAEP and TAAS stem from differences in the content of the two exams, it is necessary to categorize the items on each exam in a way that captures all relevant differences. This poses at least two

¹³ With 50-60 items per exam and over 200,000 students in each grade-year, the complete TAAS student-item level file would have over 10,000,000 observations for each grade-year alone.

¹⁴ Content and format information was not available for six multiple choice items on the NAEP exam (across both years and grades), so these six items were dropped from the analysis.

¹⁵ The summary statistics for the NAEP sample take into account the stratification and clustering associated with the NAEP sampling design and thus provide consistent and efficient estimates of population means and variances.

distinct problems. The first challenge is to determine which items should be grouped together, both within any particular exam as well as across exams.¹⁶

One approach is to use the content areas commonly reported by the test manufacturers. Norm-referenced state assessments such as the ITBS, SAT9, or CTBS4/5 generally provide separate performance measures in areas such as data analysis, measurement or algebra. Performance is reported as a mean raw score (e.g., percent of students answering the items in these areas correctly) or a scaled score. Criterion-referenced state exams report similar categories. The NAEP reports math achievement in five content strands: number sense, properties and operations; measurement; geometry and spatial sense; data analysis, statistics and probability; and algebra and functions. The primary advantage of tracking performance in these content areas is that the data is readily available. An important disadvantage of this approach, however, is that the content areas are generally determined theoretically rather than empirically. In practice, items from different areas may have more in common with each other than with items in their own content area. In addition, these content areas tend to be quite broad, so that they might mask important distinctions within area.¹⁷

An alternative approach is to track performance on individual items or groups of items. The advantage to this strategy is that it allows one to examine very specific topics. For example,

¹⁶ This might involve the same exam in different years (e.g., 1994 NAEP vs. 1998 NAEP) or two different exams in the same year (e.g., 2000 NAEP and 2000 TAAS).

¹⁷ For example, the following question is listed under objective 13 – evaluating the reasonableness of a solution: “Q: Which is a reasonable remainder when a number is divided by 5? A: 7, 6, 5, 4.” First and foremost, this problem requires that a children understand division. Another: “Q: Holly practices the piano more than Lee does. Lee practices less than Sally. John practices more than Holly. Which of the following is a reasonable conclusion? A: John practices more than Lee, Sally practices the least, Holly practices the most, John practices less than Lee.” This problem clearly requires a good deal of abstract, logical reasoning. Similarly, all of the items classified as “problem-solving using solution strategies” rely on a wide range of skills including not only arithmetic, but in some cases knowledge of area, perimeter, or measurement. On the other hand, the objectives under the operations category are fairly specific and correspond closely with more specific skill groupings, with the exception that several of the addition and subtraction problems involve decimals rather than whole numbers. However, these items have different formats that might be relevant to student performance. For example, some are straight calculation problems, some are word problems, and others have graphical or picture aides.

one can track student ability to work with number series or to identify irregular polygons. If identical items are given across years, one can track performance on the exact same question. The disadvantage of looking at individual items is that there may be significant measurement error associated with student performance on a single item, precluding reliable estimation of achievement trends.

In order to increase the reliability of this exercise, one can group similar items together and track performance trends across groups of items. The key to this approach is deciding which items should be grouped together. Items vary along a number of dimensions, allowing multiple potential groupings. Consider, for example, a math word problem that asks students to add currency units expressed in decimals. This might be classified as a basic arithmetic problem, a problem on decimals, a word problem, a problem involving currency, etc.

To create the item groupings, I first examined the standards, curriculum frameworks and test documentation for TAAS as well as NAEP. In addition, I examined all of the items administered on both exams from 1996 to 2000 in order to obtain a more holistic impression of the “type” of questions that were asked. I then examined each test item and categorized it along two dimensions – content and format.

For content, I first listed each skill that the item assessed. To provide maximum flexibility for later grouping, I defined these skills quite narrowly. In total, I defined 44 distinct skills shown on Table 3. The skills included, among others, addition, subtraction, multiplication, division, decimals, fractions, percents, units of measurement, order of operations, solving an algebraic equation, angles, volume, area, perimeter, exponential notation, bar graphs, line graphs, and pie charts and other tables or charts. Each of these skills and their definitions are described

in Appendix A. Note that many items assessed multiple skills, and therefore fell into more than one skill category.

For format, I noted a variety of factors related to the way the question appeared in the test booklet, including whether it was in the form of a word problem, whether it was accompanied by any visual displays, whether the problem contained any extraneous or potentially distracting information, and whether the problem provided a definition essential to solving the problem. Summary statistics of the format codes are shown in Table 4. A complete list of format codes and their definitions are described in Appendix B. As with skill codes, many items received multiple codes for formatting.

While I have attempted to define skills and formats in a logical manner that is consistent with the elementary school math curriculum, this process is inherently subjective. For this reason, all classifications were completed *prior* to the analysis of achievement data. Moreover, in the analysis section that follows, I experiment with a variety of possible specifications. Finally, I have made every effort to be completely transparent regarding the definition and classification of items.

A second challenge is to determine how to aggregate the information from multiple items to create a composite category score that can be compared over time. I pursue the simplest strategy, which is to consider the percentage of students who correctly answer each item correctly and then average these percentages across all items in a group. Given the stability of item type over time, this should provide a reasonable proxy for achievement trends within item grouping. However, if the composition of items in a particular grouping varies substantially over time so that more or less difficult items in this category are included in later administrations of

the exam, the analysis of percent correct might confound changes in student achievement in this skill or topic area with changes in item difficulty.¹⁸

4.3 Texas Student Performance Trends on the TAAS and the NAEP

Figures 5 and 6 compare the gains made by Texas students on the TAAS versus the NAEP during the 1990s. For fourth grade, TAAS scores rose sharply in the initial years, climbing somewhat more slowly after 1996. A similar pattern is apparent in the NAEP scores, although the pace of gains was consistently slower on the NAEP. For example, Texas fourth graders improved over 0.3 standard deviations on the TAAS between 1996 and 2000, compared with less than 0.1 standard deviations on NAEP. The differential gains are even greater for eighth graders, with TAAS scores increasing nearly 0.6 standard deviations compared with a 0.1 standard deviation on the NAEP.

There are a number of important differences between the NAEP and the TAAS that might explain the differential trends. First, the TAAS is comprised entirely of multiple-choice items whereas the NAEP includes short answers that students must complete, along with long answer questions which very open ended and graded by teachers using a rubric. Indeed, only 60 percent of the items on the NAEP exam were multiple-choice, with the remainder of the items divided about evenly between short- and long-answer open-response items.

Figures 7 and 8 compare student performance trends on the TAAS versus NAEP, focusing only on student responses to the multiple choice items from 1996 to 2000. Student performance on the TAAS continues to outpace achievement on the NAEP. Over this period, eighth grade student performance on NAEP multiple choice items remained unchanged, and

¹⁸ To the extent that one has information on the relative difficulty of items each year based, for example, on Item Response Theory (IRT) equating, were available, it would be possible to verify this assumption of stable item difficulty within item group. Such item difficulties are available for the NAEP but not the TAAS.

fourth grade performance increased only 3 percentage points. In contrast, eighth and fourth grade TAAS performance both increased by roughly 5 percentage points.

Another potentially important difference between the exams involves the use of aides. Among the multiple-choice items on the NAEP exam, 29 percent of the items required the use of calculators, 6 percent required the use of rulers or protractors and 5 percent required the use of another manipulative. None of the items on TAAS required or permitted the use of such aides.

Figures 7 and 8 also show trends of student performance on these multiple-choice items that did not involve aides. Among eighth graders, performance on these NAEP items actually declined over the period, indicating that the presence of aides cannot explain the growth in the TAAS-NAEP gap. In fourth grade, however, student performance on these items increased by 4 percentage points, quite close to the TAAS gain. This suggests that differences in the type of items across exams may be an important explanation for the gap in fourth grade.

4.4 Methodology

It is relatively straightforward to examine performance trends. I start by estimating a regression model similar to equation (1) but where the each observation reflects a student's performance on a particular item. Consider the regression equation:

$$(2) \quad y_{iesgt} = \alpha + \beta_1 Year_t + \beta_2 TAAS_e + \beta_3 TAAS_e * Year_t + \varepsilon_{iesgt}$$

where i indexes item, e indexes exam ($e = TAAS$ or $NAEP$), s indexes student, g indexes grade ($g = 4$ or 8) and t indexes year ($t = 1996$ or 2000). Here y_{iesgt} is a binary variable where the value 1 indicates a correct answer. The coefficient β_1 reflects the average annual learning gain on the TASS exam measured in terms of the percent of items answered correctly. The

coefficient on the interaction term $TAAS*Year, \beta_3$, reflects any additional annual learning gain on the TAAS, again measured in terms of the percent of items answered correctly.

Because the NAEP over-samples certain groups, all analyses with the NAEP data are conducted using the individual sampling weight (ORIGWT). Moreover, our estimates take into account the clustering and stratification inherent in the NAEP sampling design. Specifically, we assume that students are clustered within schools, and that schools are stratified by the percentage of minority students in the school and the urbanicity of the district in which the school is located.¹⁹

In addition to the inclusion of open-response items, one of the biggest differences between the two exams is that NAEP is a timed exam whereas there is no time limit imposed on students taking the TAAS. Anecdotes by Texas elementary school teachers suggest that certain students and classes were given many hours to complete the exam. It is difficult to determine how great a role this factor played in the divergence between TAAS and NAEP scores over the 1990s. However, one way to gain some insight is to examine student performance on items at the beginning versus the end of the exam. To the extent that time limits are binding, one might expect that the *difference* between student performance at the beginning versus at the end of the exam would be greater for the timed exam than the un-timed exam. Furthermore, as student motivation increases, one might expect this difference to widen. To examine this, we include indicators for item position (specifically, the quartile of the test in which the item appears) along with interactions between item position and year.

¹⁹ In practice, because the TAAS data does not come from a complex sample design, but is rather a simple random sample, we do not use strata for the TAAS data. However, we still assume that students in the TAAS data are clustered within schools since we would expect that the errors of these students would be correlated regardless of the sample design.

Another hypothesis is that the disproportionately large TAAS gains over this period were due to changes in student composition driven by exclusions from testing. While it is impossible to test the importance of selection along unobservable dimensions, we can test selection on the basis of observable demographics. To do so, we include a set of individual covariates, X , which includes indicators for race, gender, free-lunch eligibility, special education status, and limited English proficiency. We will also run separate specifications by race and gender.

Finally, in order to determine the extent to which question content or format can explain the differential achievement trends across exams, we estimate a model that incorporates information on the skills and formats relevant for each item. The full specification will resemble the following regression:

$$(3) \quad y_{iesgt} = \alpha + \beta_1 Year_t + \beta_2 TAAS_e + \beta_3 TAAS_e * Year_t + BX_s + \Delta ItemPos_i + \Phi ItemPos_i * Year_t + \Gamma C_{ie} + \Pi C_{ie} * Year_t + \varepsilon_{iesgt}$$

The inclusion of main effects for skills or formats, C , allows student performance to vary systematically across types of items. We also include a series of interactions between the skill/format categories, year and the exam indicator. We do so to allow for improvement over time to vary across category. With these additional controls, the coefficient β_3 in equation (3) now reflects the differential improvement on TAAS *within* skill or content area. By comparing β_3 in equations (3) and (2), one can determine the percentage of the TAAS effect was driven by improvement in TAAS specific items. Specifically, we will test the null hypothesis $H_0 : \beta_3 = 0$, which would indicate that the unconditional TAAS-NAEP differences can be completely explained by differential trends in certain content/format types. If we reject the null hypothesis, we will conclude that some other factors (e.g., test day effort, cheating, etc.) are driving the TAAS-NAEP differences.

Finally, it is worth noting that for the present analysis which focuses simply on individual item responses rather than overall student achievement scores, it is not necessary to explicitly take into account the matrix sampling structure of the NAEP. In order to assess a broad range of skills and topics in a limited amount of time, NAEP only administers a subset of the total cognitive items to each student in such a way that students in each school together take the full range of test items. Because this makes it impossible to compare the raw test scores across students, NAEP uses a scaling method to summarize student performance.²⁰ However, while a student's total raw score on the NAEP items he or she took is not an unbiased estimate of his or her true ability (and thus cannot be compared across students), a student's response on a particular item is indeed an unbiased estimate of the student's proficiency on that item.²¹

4.5 A Comparison of TAAS vs. NAEP within Item Type

While the Texas learning objectives and NAEP content strands provide some information to on student performance by skill, they are quite broad. If one looks at the distribution of skills across exams using the more detailed item coding scheme described above, it is clear that there are important differences between the NAEP and TAAS. Tables 3 and 4 show the distribution of skills and formats on both exams. For example, we see that a greater fraction of TAAS items

²⁰ NAEP uses a multiple imputation method to create five "plausible values" for a student's true proficiency. These plausible values are essentially random draws from the student's posterior conditional distribution, conditioning on the student's performance on the subset of items she did answer as well a multitude of background and demographic characteristics observed from the questionnaires. It is important to realize that these plausible values are not the same as individual test scores and generally not unbiased estimates of an individual student's proficiency. However, in a properly specified analysis, they can be used to provide consistent estimates of population parameters.

²¹ If the matrix sampling were conducted such that all of the really smart kids in a school, or all of the really high functioning schools, got one type of items whereas all of the really dumb kids and/or bad schools got another type of items, this would bias my estimate of the average student's proficiency on each type of item. Furthermore, if the distribution of students across items changed over time (e.g., in 2000 all of the smart kids/good schools got the items that all of the dumb kids/bad schools got in 1996, and vice versa), then estimates of changes in student ability on particular items would be biased. (In fact, if the distribution of students across items did not change, one could infer the change in the ability of certain students for certain items, but could still not infer the change in ability for the average student.) However, these possibilities seem quite unlikely.

involve basic arithmetic relative to the NAEP. For example, 14 percent of TAAS items require students to understand subtraction compared with only 9 percent of NAEP items. The differences for multiplication and division for TAAS and NAEP respectively are 19 percent vs. 13 percent and 13 percent vs. 8 percent. While 3 percent of the items on both exams require students to understand percents, 7 percent of items on the NAEP require knowledge of fractions relative to only 3 percent of TAAS items. On the other hand, 17 percent of TAAS items involve decimals compared with only 10 percent of NAEP items. The NAEP exam weighs algebra, geometry and measurement skills more heavily than the TAAS while the TAAS exam places more emphasis on estimation skills, ratios and proportions, and rate-time-distance problems relative to the NAEP.

Table 4 reveals several somewhat unexpected findings with regard to format differences. Perhaps most interestingly, TAAS multiple-choice items are nearly 50 percent more likely to be written as a word problem relative to NAEP multiple-choice items. On the other hand, NAEP items are more than 50 percent more likely than TAAS items to contain a pictorial aide in the context of the question. TAAS items are also much more likely to require students to merely identify a strategy for solving a problem rather than actually solving the problem. Perhaps the most obvious difference between the two exams, however, involves the use of what are referred to as “manipulatives,” which includes calculators, protractors, rulers, etc. TAAS neither requires nor allows any of these aides. In contrast, roughly 29 percent of NAEP multiple-choice items permit the use of a calculator, 5 percent require the use of rulers or protractors and 6 percent involve the use of some other manipulative.

But do these differences in content and format explain the differences in student achievement trends between 1996 and 2000 on the two exams? To answer this question, I

estimate specifications based on equation (3). Table 5 presents results from preliminary regressions. In column 1, we see that across both exams, students made some improvement from 1996 to 2000, and that eighth graders scored roughly 5 percentage points lower than fourth graders. Perhaps most importantly, we see the stark difference in difficulty across exams that we illustrated earlier. Across both years and grades, students were roughly 20 percentage points more likely to answer an item on the TAAS correctly relative to the NAEP. Given the baseline percent correct of roughly 50 percent on the NAEP, the relative performance gap between the TAAS and NAEP was approximately 40 percent over this period.

Column 2 includes an interaction term to capture the change in the TAAS-NAEP gap from 1996 to 2000. The coefficient indicates that the gap increased by 4.5 percentage points (roughly 25 percent) over this period. Column 3 includes controls for student demographics. Not surprisingly, we find that poor children and children with Limited English proficiency or special education placements score significantly lower than their peers. Similarly, Black and Hispanic children score significantly lower than white or Asian children. While the inclusion of these demographic variables increases the explanatory power of the models, these variables do not change the estimate of the TAAS-NAEP gap in 1996 or the change in the gap from 1996 to 2000. This suggests that changes in student composition, at least along easily observable dimensions, do not explain the growth in the TAAS-NAEP gap.

Column 3 also includes indicators for whether the item was in the second, third or fourth quartile of its respective exam. Students do best on items in the first quarter of an exam and worse on items in the third quarter. Column 4 includes interactions between item position and year. In this specification, we see that across both exams between 1996 and 2000 students improved more on items in at the end of the exams relative to items at the beginning of the

exams. Note, however, that the coefficient on the TAAS*Year 2000 interaction does not change appreciably from column 3 to column 4, indicating that this “end of exam” improvement was not disproportionately focused on the TAAS exam. This suggests that increases in student test-taking persistence may not explain the growth in the TAAS-NAEP gap from 1996 to 2000.

Columns 5-6 present the results separately for grades four and eight. Importantly, we see that the TAAS-NAEP gap grew by 5.6 percentage points or 30 percent among eighth graders compared with only 2.6 percentage points or 11 percent among fourth graders. This suggests that test score inflation in math over this period may have been more prevalent in the higher grades.

Table 6 presents estimates specifications that include controls for the skills assessed and the formats used on both exams. Column 1 reproduces the raw change in the TAAS-NAEP from 1996 to 2000 and column 2, which controls for student demographics and item position, replicates the estimates described above (from Table 5). Columns 3-6 control for the skills assessed in the items as well as the formats in which the items are presented.

As we saw in Tables 3 and 4, the items on the two exams assess 44 separate skills and have 17 specific formatting characteristics. Moreover, many items assess multiple skills and have multiple format characteristics, so that the total number of combinations of skills and formats is extremely large. However, we can also see from Tables 5 and 6 that many of the skills and formats are very rare, and in many cases only involve a single item on a single exam. For the sake of parsimony therefore, I began by considering only 14 skill groupings and 8 format groupings. Specifically, I created 14 indicator variables indicating whether the item assessed the following 14 skills: 1) addition, 2) subtraction, 3) multiplication, 4) division, 5) probability, 6) spatial reasoning, 7) percents, 8) decimals, 9) fractions, 10) Greater than/less than, 11) any type

of graph (including Cartesian, line, pie, bar or other), 12) algebra (either writing or solving algebra equations), 13) any type of geometry (including area, perimeter, volume, Pythagorean Theorem, angles, or shapes and figures), and 14) time. Recall that items which assess multiple skills may have a value of 1 for more than one of these indicator variables. For items that did not assess any of the listed skills, all 14 variables take on a value of zero. In a similar fashion, I created 8 indicator variables indicating whether the item contained the following formatting characteristics: 1) requires use of rulers, protractors or other manipulative, 2) requires use of calculators, 3) word problem, 4) includes picture aide, 5) requires solution strategy only, 6) question asked in the negative, 7) five possible choices are provided rather than four, and 8) problem contains extraneous information.

The specification shown in column 3 includes 14 binary skill indicators along with 14 interaction terms between each skill and the year 2000. The inclusion of these interaction terms accounts for changes in student performance in specific skills over time. If we believe, for example, that the TAAS gains over this period were driven by improvement in skills that were emphasized disproportionately on the TAAS exam, we would expect the inclusion of these controls to diminish the growth in the TAAS-NAEP gap. In fact, we see that the gap increases slightly for both fourth and eighth graders. Column 4 includes controls for the 8 format indicators and interactions between these variables and the year 2000. The inclusion of these format controls does not change the gap for eighth graders, but dramatically reduces the gap among fourth graders. Indeed, the results in column 4 suggest that all of the disproportionate TAAS gains in fourth grade came from items with formatting that was unique to the TAAS exam. If one accounts for skills and formats simultaneously in column 5, the gap remains essentially unchanged relative to column 2.

While the model in column 5 controls for skills and formats individually, it does not allow for interactions among different skills or formats. For example, if teachers in Texas drilled students in decimals over this period and students made substantial gains in all types of problems involving decimals, the specification above would not allow these gains to influence the TAAS-NAEP gap. The model would not, however, account for a more nuanced situation in which, for example, teachers drilled students on decimal addition but not other problems involving decimals. In column 6 will include controls that capture the interactions between skill and format combinations. Specifically, we include 145 indicator variables for the most common skill*format combinations in the data. This essentially means that we are looking at changes in student performance over time *within* very specific types of problems. For example, these controls allow us to distinguish between decimal addition problems that are presented in the context of a word problem from decimal addition problems that are presented as a simple calculation problem, or between problems involving multiplication of fraction that stem from reading a bar graph from problems involving multiplication of fractions that stem from reading a pie chart.

The results in column 6 suggest that skill and format explain virtually all of the growth in the TAAS-NAEP gap for fourth graders, and roughly one-third of the growth among eighth graders. As one seeks to interpret the estimates in Table 6, however, it is important to keep in mind that the inclusion of skill (format) controls means that the change in the TAAS-NAEP gap is only being estimated from items which have skills (formats) that are common to both exams. Items with skills (formats) unique to either exam do not contribute to the estimate of the coefficients shown in Table 6. For example, the estimates in column 4 do not use variation in student performance on items in which the correct response is not provided or that have five

possible item choices instead of four since no items on the NAEP exam have these features. Similarly, the estimates in column 4 do not utilize variation in items that involve calculation questions that are displayed vertically since such items were unique to the NAEP.²²

If there is considerable overlap on the two exams, this is not particularly problematic. However, as one defines skill and format categories more precisely, the number of categories that contain items from both exams shrinks considerably. Consider, for example, the specification in column 6 which includes indicators for 145 separate categories that capture very specific combinations of skills and formats. While this model is able to capture very fine-grained distinctions between items, there are only 6 or 7 categories that contain items from both the TAAS and the NAEP. Hence, the estimates in column 6 reflect the differential student improvement on the TAAS for a relatively small subset of items.

For this reason, it is useful to consider the specifications in columns 5 as well as 6 in assessing the role that skill and format plays in explaining differential TAAS improvement over this period. Overall, the data suggest that skill and content differences across the two exams may play a large role in explaining the differential TAAS growth among fourth graders but cannot fully explain this phenomenon among eighth graders.

Table 7 presents results separately by race and gender. Several interesting facts stand out. First, the TAAS-NAEP gap grew more among Blacks and Hispanics relative to whites and more among boys relative to girls, suggesting that test score inflation was more pronounced for the former. Second, the importance of skill and format controls appears to differ more across grades than across race or gender. Specifically, the inclusion of these controls reduces the

²² This discussion is not strictly correct with regard to the estimates in columns 3-7 since the skill and format indicators are not mutually exclusive, though it is correct for the estimates in column 8 which include mutually exclusive indicators. The intuition, however, is useful in understanding the limitations of the estimates presented here.

growth in the gap substantially for all groups of fourth graders, but has little impact for any of the eighth grade subgroups. While the estimates are not particularly precise, there is an indication that even the most extensive set of skill and format controls does not substantially impact the gap for Black fourth graders. Perhaps the most interesting finding is that the inclusion of individual skill and format controls actually *increases* the growth in the gap for eighth graders. While this pattern was evident for the full sample in Table 6 (compare columns 2 and 5), the differences are much more dramatic when one looks separately by race and gender. This suggests that there are important interactions between race (gender) and student achievement across items. Even the most extensive set of combined skill-format controls does not reduce the gap relative to simply controlling for demographics and item position.

The results in Table 7 reinforce the finding from Table 6 that skill and format differences across exams might explain the disproportionate improvement in the TAAS for fourth graders, but cannot explain the similar, but even more pronounced, pattern among eighth graders. This suggests that other factors such as student effort (not related to the persistence capture by item position), exclusion on the basis of unobservable student characteristics, or cheating might be responsible for the differential improvement on the TAAS by older students over this period.

5. Conclusions

This paper explores the extent and potential causes of recent test score inflation in elementary school. Examining four states with quite different state testing regimes, I find that annual student achievement gains on state assessments have regularly outpaced performance gains on the NAEP. In a detailed study of achievement trends in Texas, I find that the differential TAAS improvement cannot be explained by changes in the demographic composition of the test-

takers, or by several important differences across the exams, including (a) the fact that the NAEP includes open-response items whereas the TAAS only includes multiple-choice items, (b) the fact that many of the multiple-choice items on the NAEP require and/or permit the use of calculators, rulers, protractors or other manipulative such blocks or fraction bars while none of the items on the TAAS do so, or (c) the fact that the TAAS is an un-timed exam and the NAEP is a timed exam. However, I do find that skill and format differences across exams might explain the disproportionate improvement in the TAAS for fourth graders, but cannot explain the trends among eighth graders. These results suggest that greater attention should be paid to the generalizability of student achievement gains under high-stakes accountability programs such as NCLB.

References

- Amrein, Audrey L. and Berliner, David C. "High-Stakes Testing, Uncertainty, and Student Learning." *Education Policy Analysis Archives* 10(18).
- Baker, George, "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, C (1992), 598-614.
- Cannell, J.J. 1987. *Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States are Above the National Average*. Daniels, W.V.: Friends for Education.
- Carnoy, Martin and Susanna Loeb (2002). "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." Working Paper, Stanford University School of Education.
- Cullen, Julie Berry and Reback, Randall (2002). "Tinkering Toward Accolades: School Gaming under a Performance Accountability System." Working paper, University of Michigan.
- Dee, Thomas S. (2002). "Standards and Student Outcomes: Lessons from the 'First Wave' of Education Reform." Working paper.
- Deere, D. and W. Strayer (2001). "Putting Schools to the Test: School Accountability, Incentives and Behavior." Working paper. Department of Economics, Texas A&M University.
- Elmore, Richard (2002). "Unwarranted Intrusion." *Education Next*. Spring, 31-36.
- Figlio, David N. and Joshua Winicki (2001). "Food for Thought: The Effects of School Accountability Plans on School Nutrition." Working Paper, University of Florida.
- Figlio, David N. (2002). "Testing, Crime and Punishment." Working Paper, University of Florida.
- Figlio, David N. and Lawrence S. Getzler (2002). "Accountability, Ability and Disability: Gaming the System?" Working Paper, University of Florida.
- Glaeser, Edward, and Andrei Shleifer, "A Reason for Quantity Regulation," *American Economic Review*, XCI (2001), 431-435.
- Greene, Jay P., Winters, Marcus A. and Greg Forster (2003). "Testing High Stakes Tests: Can We Believe the Results of Accountability Tests?" Report #33, Center for Civic Innovation, Manhattan Institute, New York City.
- Grissmer, D.W. et. al. (2000). *Improving Student Achievement: What NAEP Test Scores Tell Us*. MR-924-EDU. Santa Monica: RAND Corporation.

- Hanushek, Eric A. and Margaret E. Raymond (2002). "Improving Educational Quality: How Best to Evaluate Our Schools?" Paper prepared for a conference by the Federal Reserve Bank of Boston, June 19-21, 2002. Conference titled "Education in the 21st Century: Meeting the Challenge of a Changing World."
- Heubert, J. P. and R. M. Hauser, Eds. (1999). *High Stakes: Testing for Tracking, Promotion and Graduation*. Washington, D.C., National Academy Press.
- Holmstrom, B. and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design. *Journal of Law, Economics and Organization*. 7(Spring), 24-51.
- Jacob, B. A. (2001a). "Getting Tough? The Impact of Mandatory High School Graduation Exams on Student Outcomes." *Educational Evaluation and Policy Analysis*. 23(2): 99-121.
- Jacob, B. (2002). Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools. National Bureau of Economic Research, Working Paper #8968.
- Jacob, Brian A. and Levitt, Steven D. (Forthcoming). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*.
- Klein, S. P., L. S. Hamilton, et al. (2000). What Do Test Scores in Texas Tell Us? Santa Monica, CA, RAND.
- Kolen, Michael J. (1995). *Test Equating: Methods and Practice*. New York: Springer-Verlag.
- Kolker, Claudia. (1999). "Texas Offers Hard Lessons on School Accountability." *Los Angeles Times*, April 14, 1999.
- Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. In E. Hanushek, J. Heckman, and D. Neal (Eds), *Designing Incentives to Promote Human Capital*. Special issue of *The Journal of Human Resources*, 38(4), Fall.
- Koretz, D., McCaffrey, D., and Hamilton, L. (2001). *Toward a Framework for Validating Gains Under High-Stakes Conditions*. CSE Technical Report. Los Angeles: Center for the Study of Evaluation, University of California (in press).
- Koretz, D.M., and S.I. Barron. 1998. "The Validity of Gains on the Kentucky Instructional Results Information System." (KIRIS). Santa Monica: RAND.
- Koretz, D.M., R.L. Linn, S.B. Dunbar, and L.A. Shepard. 1991. "The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests." In R.L. Linn (chair), *The Effects of High Stakes Testing*, symposium presented at the annual meetings

of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April.

Linn, R.L. 2000. "Assessment and accountability." *Educational Researcher*, 29(2): 4-16.

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

Linn, R.L., M.E. Graue, and N.M. Sanders. 1990. "Comparing State and District Test Results to National Norms: The Validity of the Claims That `Everyone Is Above Average.'" *Educational Measurement: Issues and Practice*, 9(3): 5-14.

Lillard, Dean R. and Philip P. DeCicca (2001). "Higher Standards, More Dropouts? Evidence Within and Across Time." *Economics of Education Review* 20: 459-473.

Loughran, Regina, and Thomas Comiskey (1999). "Cheating the Children: Educator Misconduct on Standardized Tests." Report of the City of New York Special Commissioner of Investigation for the New York City School District, December.

Marcus, John. (2000). "Faking the Grade." *Boston Magazine*, February.

May, Meredith. (1999). "State Fears Cheating by Teachers." *San Francisco Chronicle*, October.

McNeil, L. M. and Valenzuela, A. (2001). "The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric." In Orfield, G. and Kornhaber, M.L. (Eds.) *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*. New York: The Century Foundation Press.

Mislevy, R. J., Johnson, E. G. and Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics* 17(2): 131-154.

Raudenbush, Stephen W. and Anthony S. Bryk (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd Edition. Sage Publications: London.

Raudenbush, Stephen, Bryk, Anthony, Cheong, Yuk Fai and Richard Congdon (2000). *HLM5: Hierarchical Linear and Nonlinear Modeling*. Scientific Software International.

Richards, Craig E. and Sheu, Tian Ming (1992). The South Carolina School Incentive Reward Program: A Policy Analysis. *Economics of Education Review* 11(1): 71-86.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley and Sons.

Shepard, L.A. 1990. "Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test?" *Educational Measurement: Issues and Practice*, 9(3): 15-22.

Shepard, L.A. 1988. "The Harm of Measurement-driven Instruction." Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. (April).

Shepard, L.A. and Dougherty, K.C. (1991). *Effects of High-Stakes testing on Instruction*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC ED 337 468).

Tepper, Robin, Susan Stone and Melissa Roderick. 2002. "Ending Social Promotion: The Response of Teachers and Students." Consortium on Chicago School Research.

Appendix A – Item Skill Classifications

Definitions used to categorize the skills assessed in items on both the Texas Assessment of Academic Skills (TAAS) and the National Assessment of Educational Progress (NAEP) mathematics exams in grades 4 and 8 from 1996 to 2000. Items can assess multiple skills, in which case they will have more than one skill code.

Skill 1 – Addition

One of the four arithmetic operation codes, questions coded 1 are ones that require addition to solve. These may take the form of word problems or calculations. If the question involves addition of decimals or fractions, an additional code (14, 15) will be added. The code for addition is included in all problems where the operation is required to solve correctly unless otherwise indicated in the definition of the primary skill.

Skill 2 – Subtraction

One of the four arithmetic operation codes, questions coded 2 are ones that require subtraction to solve. These may take the form of word problems or calculation. If the question involves subtraction of decimals or fractions, an additional code (14, 15) will be added. The code for subtraction is included in all problems where the operation is required to solve correctly unless otherwise noted in the definition of the primary skill.

Skill 3 – Multiplication

One of the four arithmetic operation codes, questions coded 3 are ones that require multiplication to solve. These may take the form of word problems or calculations. If the question involves multiplication of decimals or fractions, an additional code (14, 15) will be added. The code for multiplication is included in all problems where the operation is required to solve correctly unless otherwise noted in the definition of the primary skill. Code 3 includes knowledge that any number multiplied by zero is zero, as needed by one question on NAEP.

Skill 4 – Division

One of the four arithmetic operation codes, questions coded 4 are ones that require division to solve. These can vary in difficulty. A problem does not have to involve working with remainders to be coded a 4, although several NAEP division items where calculator use is permitted do involve working with remainders, and some items on TAAS ask students to identify a reasonable remainder from a set of choices (see 1996 Grade 4 TAAS #24). These questions are not distinguished in any way from other division problems. If the question involves division with decimals or fractions, codes (14, 15) will be added. The code for division is included in all problems where the operation is required to solve correctly unless otherwise noted in the definition of the primary skill.

Skill 5 – Probability

Questions coded 5 are ones that require an understanding of the concept of probability or the likelihood of a particular outcome in a given set of circumstances. Questions coded with Skill 5 will not be accompanied by arithmetic codes or the fraction code, by definition. This was done because probability items often require the ability to sum a series of numbers correctly and the knowledge of writing a probability in the form of a fraction (the total number of possible outcomes in the denominator and the possible number of a single outcome in the numerator).

To select the correct answer choice, students may need to reduce a probability fraction they have calculated or written. Grade 4 TAAS questions test knowledge of probability to a lesser degree than Grade 8 TAAS or NAEP. For example, these items ask students to identify the “*most likely*” outcome in a given circumstance as opposed to a Grade 8 item which asks students what the probability is of a certain outcome in a given situation (see 2000 Grade 8 TAAS #27). It should be noted that on the Grade 8 TAAS exams probabilities are written in three ways: a) x/y , b) x out of y , c) x in y . Also note that in NAEP, probabilities answers are generally written as x out of y .

Skill 6 – Spatial Sense or Reasoning/Perspective

To answer questions coded 6, students must have an understanding of perspective (e.g., how a picture would look from a different angle or position, such as from an aerial perspective - see 2000 Grade 8 TAAS #14) or employ the use of spatial reasoning to determine how many of a small object might fit into a larger object (e.g., small cubes in a bigger cube) or determine what 3-dimensional figure a 2-dimensional figure could be folded into. Questions coded skill 6 will not be combined with other skill codes.

***Not used on Grade 4 TAAS exams.*

Skill 7 – Percents

Questions coded 7 are ones that require knowledge of percentages. By definition, knowledge of percentages includes the ability to convert decimals to percents and vice-versa. These questions may require students to calculate percents from given information, identify the correct operation or number sentence to solve for a percent or calculate the number of units included in a given percentage of a whole. Percent code includes the understanding that a circle (or pie) represents 100% and that it can be sectioned to show how percents are distributed among different sets of conditions, choices, etc. Skill 7 will not be accompanied by arithmetic operation codes—it is understood that to answer these questions, addition, subtraction, multiplication or division may be needed. However, other skill codes will be used, if necessary, including decimals, fractions, estimation, perimeter, etc. It should be noted that, on TAAS exams, when asked to identify the correct calculation to find a percent, answer choices are listed in one of the three following formats: a) $325/1700 = x/100$, b) $6/24 \times 100$ or c) $(6)(100)/15$ (see 1999 Grade 8 TAAS #41 compared with 1997 Grade 8 TAAS #25).

*** Not used on Grade 4 TAAS exams.*

Skill 8 – Rate/Time/Distance

Questions coded 8 are ones that require knowledge of Rate/Time/Distance problems (commonly recognized as a specific type of problem taught as a unit in middle-school math classes). R/T/D/ problems require an understanding of constant rate of change in a given time period (e.g., miles traveled per hour, inflation per minute, etc.) and almost always require division or multiplication to solve. Students may be given any two of the three pieces of information (i.e., rate, time or distance) and asked to solve for the third value (see 1996 Grade 8 TAAS #53). Skill 8 will not be accompanied by arithmetic codes. However, other skill codes will be used, if necessary, including decimals, fractions, estimation, etc.

***Not used in NAEP or Grade 4 TAAS exams.*

Skill 9 – Ratio/Proportion

Questions coded 9 are ones that require an understanding of ratios. This may include an understanding of the concept of proportion (e.g., a photo is enlarged, therefore the perimeter of the enlargement is proportional to the perimeter of the original photo). Skill 9 will not be accompanied by arithmetic codes. However, other codes—including those for decimals, fractions, estimation, perimeter and area – will be added as needed. Typical skill 9 questions include surveys (e.g., if 15 out of 100 people prefer Cereal A, how many people would prefer Cereal A if we ask 1000 people?), writing ratios of wins to losses (e.g. XYZ school’s soccer team has a record of 10 wins and 5 losses, what is the ratio of wins to losses?), and recognizing the operation needed to solve for a missing value in a calculation using ratios (e.g., if $\frac{3}{5} = \frac{x}{20}$, what operation would you use to find the value of x?). Skill 9 questions use the following three notations for ratios/proportions: a) x out of y, b) $\frac{x}{y}$ and c) x to y.

It should be noted that although use of the word “rate” in some of these may seem similar to skill 8, we distinguish between the two codes by defining 8 as a code for questions involving a rate of change over time or distance, while skill 9 questions involve a specific ratio at a moment in time. For example, consider the following two problems: 1) “An experienced diver descends from the surface at a rate of 4 meters per minute, what is her location at the end of 10 minutes?” – Skill 8 Rate/Time/Distance 1996 Grade 8 TAAS. 2) “A certain medication for dogs is prescribed according to body weight. A 15 lb dog requires 45 mg of medication. At the same rate, how much medication would a 20 lb dog require?” – Skill 9 Ratio/Proportion 2000 Grade 8 TAAS

Skill 10 – Measurement Conversion

Questions coded 10 are ones that require a conversion of a given value in specified units to the equivalent value in a different unit (e.g., kilograms to pounds). By definition, these questions require use of information given in the question or in a supplied table that details the conversion formula. Code 10 is not accompanied by arithmetic codes, although these questions often involve multiplication or division to solve. However, if the question involves decimals, fractions, or knowledge of another skill, the corresponding codes will be used. For example, the following type of question would receive a code of 10 as well as a code of 11 (indicating estimation): “One kilogram is about 2.2 pounds. If Sara weighs 94 pounds, then she must

weigh...?” Answer choices are written as estimates (e.g. less than x or between x and y) 1996 Grade 8 TAAS # 40”.

***Not used on Grade 4 TAAS exams.*

Skill 11 – Estimation

Questions coded 11 are ones that involve estimation. Typically, these are word problems that ask students to choose a reasonable estimate for the question asked. Answer choices are given either as number values or as a range (i.e. “between x and y” or “less than x”). To answer these questions correctly, students must have an understanding of rounding numbers to the nearest multiple of ten, or the nearest whole number for quick calculation. These items may also give students a range of values for an item and require that they estimate a total. For example, person A bought 7 books each cost between \$3-\$8, what is the reasonable estimate of the total cost? Skill code 11 will be combined with other skill codes including, the arithmetic operation codes needed to solve the problem and with a code of 14 or 15 if the problem involves decimals or fractions (see 1997 Grade 8 TAAS #27). It should be noted that NAEP questions involving estimation are unlike the TAAS items described above. There are only 5 questions that have skill code of 11, and it is listed as the 2nd or 3rd skill on each of them. For example, these items may require students to estimate the value of points on a line graph before solving or to estimate the value of a point on a number line between two whole numbers.

Skill 12 – Logical Reasoning

Questions coded skill 12 are ones that test logical reasoning. While nearly all mathematical problems, particularly word problems, could be considered to involve logical reasoning, skill 12 is meant to reflect a very specific type of problem in which students are given a set of circumstances or a set of relationships and asked to determine which conclusions are reasonable, logical or possible. For example, “Jacques, Tanika and Mike ran for class president. Tanika received 50 more votes than Mike. Jacques received twice as many votes as Mike. There was a total of 270 votes. Which of the following is a reasonable conclusion about the votes received?” (1999 Grade 8 TAAS #28) The answer choices follow this pattern: “Tanika received fewer votes than Mike; Mike received the least number of votes, etc.”. Skill 12 is only combined with other skill codes when the information provided is given in the form of a graph, diagram or other table (i.e. skill codes 21, 22, 23 and 35 may be added).

Skill 13 – Patterns, Recognizing and Replicating

Questions coded 13 are ones that test the ability to recognize what pattern is shown in the given question and to identify what the next value in the pattern should be or what value in a pattern has been left out. These patterns may be numerical or pictorial in nature. For example, consider the following: 1) a TAAS Grade 4 question that lists the following dates (1840, 1845, 1850, 1860, 1865, 1870) and asks which year was left out versus 2) a NAEP question that gives four pictures of a circle with shaded wedges, each figure has been rotated, students must select the circle that is next in the rotation pattern. Skill 13 will be combined with other skill codes, as needed.

Skill 14 – Decimals

Code 14 indicates the presence of numbers with decimals and the need to work with them to solve the question correctly. This is combined with all other skill codes. Items with decimals may involve the following: 1) word problems that require arithmetic with decimals to solve (see 1997 Grade 8 TAAS #59); 2) estimation of a decimal given a point on a number line between two whole numbers (see 1996 Grade 4 TAAS #17); 3) determining the decimal represented by a partially shaded figure (see 1999 Grade 4 TAAS #6); or 4) putting numbers with decimals in order from least to greatest. It should be noted that many word problems involving calculations with decimals deal with dollars and cents.

Skill 15 – Fractions

Code 15 indicates the presence of fractions and the need to work with them to solve the question correctly. It is combined with all other codes unless otherwise noted. Items with fractions may involve the following: 1) arithmetic problems with fractions (see 1997 Grade 8 TAAS #54); 2) identifying fractions that are greater than or less than other fractions (see 2000 Grade 4 TAAS #8); 3) determining the fraction represented by a partially shaded figure (see 2000 Grade 4 TAAS #10) or 4) identifying equivalent fractions (given two shaded figures as aids) (see 1997 Grade 4 TAAS #10).

Skill 16 – Negative Numbers

A skill code of 16 indicates the presence of negative numbers in a question. It is often used in conjunction with the skill codes 1-4 (for arithmetic operations) or skill 17 (greater than/less than). By definition, this is not a primary skill and therefore is listed as the second or third skill. It can be assumed that items with a skill code of 16 require students to work with negative numbers in problems of a type defined by the other skill codes present.

***Not used on Grade 4 TAAS.*

Skill 17 – “Greater than/Less than”

Questions coded 17 are ones that require an understanding of numbers and their relative values (i.e. are they greater than or less than another number or numbers). Typical questions in this category will ask students to select the answer choice that shows a set of numbers in order from least to greatest or to identify a set of numbers shown on a number line (e.g. “What number line shows the graph of the whole numbers greater than 4 and less than 7?” 1998 Grade 4 TAAS #10). By definition, this code assumes students know the definition of a whole number and understand place value. Skill 17 may be combined with a limited set of other skill codes including the following; 14 or 15 for decimals and fractions, 16 for negative and positive numbers, 21-24 if the information must be read from a chart or graph, 34 when comparing the size of angles, and the arithmetic operation codes, if the solution involves a calculation. For example, 1999 Grade 8 TAAS #26 reads “The chart shows the number of gold, silver and bronze medals won by countries with at least 10 medals in the 1994 Winter Olympics. If these 7 countries were ranked in order by the total number of medals won, which 3 countries would be ranked first, second and third?”.

Skill 18 – Statistics

Questions coded 18 assume knowledge of basic descriptive statistics (mean, median and mode) and the ability to differentiate amongst them and calculate one from a given set of data. Code 18 is not accompanied by arithmetic codes. By definition, these questions require the ability to calculate means or averages by correctly summing a series of values and dividing by the number of observations. When necessary, code 14 or 15 will be added to note the presence of or need to calculate numbers with decimals or fractions. It should be noted that some questions ask students to calculate the “mean” while others simply ask for the “average”. If accompanied by code 11, estimation, the item requires estimation of the reasonable average from the numbers given. It should be noted that on TAAS exams, when skill code 18 is accompanied by a P format code, the data given is displayed in a table.

*** Not used on Grade 4 TAAS exams.*

Skill 19 - Rounding

Questions coded 19 are ones that ask students to round a given number to the nearest specified number place (i.e. hundred, million, tenth, etc.). It should be noted that on NAEP this may include rounding to the nearest dollar or hour, as well. This skill code is typically used alone and is only ever combined with skill code 14 to indicate that rounding numbers with decimals is required.

Skill 20 – Cartesian Graphs

Questions coded 20 test an understanding of and ability to work with Cartesian graphs. This knowledge includes correctly naming coordinates, the ability to plot points using given coordinates, understanding the x- and y-axes and four quadrant structure of a Cartesian graph, and knowing the basic equation of a line parallel to the x- or y- axis. For example, typical questions may ask students to name the coordinates of the intersection between two lines, to select the axis of symmetry of a given figure or to identify the equation of a given line.

Questions that ask students to identify the axis of symmetry for a given figure are also coded 81 Geometry (see 1999 Grade 8 TAAS #12).

***Not used on Grade 4 TAAS exams.*

Skill 21 – Line Graphs

Questions coded 21 involve the interpretation of a line graph. By definition, this assumes students can read data or information from a line graph, using axis labels to help them. Typical questions will require students to read one or more points on the graph to obtain the information or values needed to answer the question correctly. This usually involves reading the information from the graph and performing operations to solve the stated question. For example, consider question #42 from the Grade 8 TAAS exam in 1996 where a graph depicts the temperature of two patients at a given time, and the question reads “Patient #2 was given 30 grains of

medication every time his temperature rose at least 2° in an hour. How many grains of medication did he receive on July 5 between the hours of 8 A.M. and 2 P.M.?” If calculations are needed to solve, then arithmetic or other skill codes will be added, these are predominantly addition and subtraction.

Skill 22 – Pie Charts

Questions coded 22 test the ability to read information depicted on a pie chart (i.e., circle graph). This ability hinges on knowing that pie charts represent a whole that is divided into subcategories or sections. Usually this code will be combined with skill code 7 - Percents, as pie charts are frequently used to show the percentages of different categories that make up a 100% sample or whole. By definition, these questions will not be combined with arithmetic codes, – the ability to add or subtract percents is implicit. Skill 22 may be combined with skills 14 or 15 if the question involves fractions or decimals (see 1996 Grade 8 TAAS #21). A typical skill 22 question will ask the student to identify the sections of the pie chart that meet a set of criteria (e.g., add up to $\frac{1}{2}$ or 50%) or to select the pie chart that best represents the information provided. For example, consider question # 10 on the Grade 8 TAAS exam in 2000 “A city’s curbside recycling program reported the percents by weight of the materials collected and processed for 1 year in the following table. Which circle graph best represents the recycling data?”

***Not used on Grade 4 TAAS exams*

Skill 23 – Bar Graphs

Questions coded 23 are ones that require an understanding of and ability to read information depicted on a bar graph. These items typically involve comparing information across bars on a given graph, combining information that meets certain criteria and performing calculations to answer the stated question. This skill code will be combined, when necessary, with other codes including the arithmetic operation codes.

Skill 24 – Other Tables or Charts

Questions given a skill code of 24 involve information that is organized into a chart or table. These questions require students to read information correctly from the table, combine values that fit a given set of criteria and perform arithmetic operations to answer a stated question. For example, question 3 on the 1996 Grade 8 TAAS exam reads “The chart shows the results of a survey on popular ice-cream flavors. How many flavors were chosen by fewer than 4 students?”. When necessary, this code is combined with all other skill codes.

Skill 25 – Exponents (general)

Questions coded 25 require an understanding of exponents. For example, 1997 Grade 8 TAAS #1 asks students to select from the given answer choices another way to write $2 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot a \cdot a$ and has a correct answer of $2^3 \cdot 3^2 \cdot a^2$. This skill code is most often used alone. However, when listed as the second or third skill it indicates that the question requires students to use square roots.

***Not used on Grade 4 TAAS exams.*

Skill 26 – Exponential Notation

Questions coded 26 require knowledge of exponential or scientific notation. These items will provide a number written in scientific notation and ask students to select another way to express this number. For example, consider question #9 on the Grade 8 TAAS exam in 1999 “The speed of light through glass is 1.22×10^5 miles per second. Which is another way to express this measure?”, which has a correct answer of “122,000 miles per second”. Alternatively, an item may write out a number and ask the student to identify how it would be written in scientific notation (see NAEP). It should be noted that when combined with code 16 (negative/positive numbers), it indicates that the scientific notation is raising a number to a negative power (e.g. 1.0×10^{-3}).

***Not used on Grade 4 TAAS exams.*

Skill 27- Writing Algebraic Expressions (Equations and Inequalities)

Questions coded 27 involve identifying the correct algebraic equation or inequality for the information provided. Typical questions are word problems that provide some information and ask students to identify the correct expression to use in order to find the value for x (a missing piece of information). For example, item #41 on the 1996 Grade 8 TAAS exam reads, “Two angles of a triangle have the same measure, and the third angle measures 68 degrees. Which equation could be used to find M , the measure of each of the 2 congruent angles?” with a correct answer of “ $2M + 68 = 180$ ”. Skill code 27 is used alone and is not combined with other skill codes. By definition, it requires knowledge of how to apply arithmetic operations to solve a word problem and how to write an expression for this solution strategy using both numbers and variables. Questions coded 27 will also be given a format code of S, since these, like some other items, involve identification of the correct solution strategy.

***Not used on Grade 4 TAAS exams.*

Skill 28 – Solving and Understanding Algebraic Expressions

Questions coded skill 28 are ones that require students to understand and solve algebraic equations and inequalities. Typical questions involve solving for a variable in a given expression, identifying all of the numbers that make a given expression true, graphing a given inequality on a number line or describing what would happen to the value of a variable on one side of an equation if the variable on the opposite side is changed (i.e. doubled, increased by a certain amount, etc). Unlike skill code 27, these items may be combined with other skill codes as necessary. For example if the question requires calculation of the value of a variable in an algebraic expression, it will be coded with the appropriate arithmetic operations codes as in #22 on the 1999 Grade 8 TAAS exam.

***Not used on Grade 4 TAAS exams.*

Skill 29 – Area

Questions coded 30 fall into one of two categories. The first type of question asks students to estimate or calculate the area represented by the shaded portion of a figure where the figure is drawn on a grid so that the student can calculate the area by simply counting the number of

square units that are shaded (see 1999 Grade 8 TAAS #44). The second type of question requires students to calculate the area of a given shape or object by reading the measurements of the sides of the figure and applying the correct formula (see 1996 Grade 8 TAAS #37). When calculation is required, the appropriate skill codes, including arithmetic, will be added. It should be noted that all skill 30 items on TAAS provide a picture or visual aid while NAEP may, for example, simply ask students to calculate the area of a square with a side of a given length.

Skill 30 – Perimeter

Questions coded 31 require students to understand the concept/definition of perimeter. Typical questions either ask students to calculate the perimeter of a figure given information on the length of one or more sides or, conversely, ask students to calculate the length of a side given information on the perimeter of the figure. By definition, these questions involve addition and subtraction and will not be accompanied by arithmetic codes. Where necessary or appropriate, other skill codes will be used.

Skill 31 – Volume

Questions coded 32 require students to understand the concept of and to calculate volume. Typical questions will either provide the dimensions of an object and ask for the volume or provide the volume of an object with some information about its dimensions and ask for the value of an unlabeled dimension (see 2000 Grade 8 TAAS #16). By definition, these questions involve arithmetic operations (usually multiplication and division) and will not be accompanied by these skill codes. If necessary, a code for decimals or fractions will be given. It should be noted that, the formulas needed to calculate volume are provided on a chart (on TAAS) or in the question (on NAEP). Volume questions on these exams include volume of a prism, sphere and cylinder.

Only 1 question has code 32 on NAEP, and it concerns the volume of a sphere.

***Not used on Grade 4 TAAS exams.*

Skill 32 – Pythagorean Theorem

Questions coded 33 involve use of the Pythagorean Theorem. To answer these items correctly, students must calculate the value of an unlabeled side of a right triangle using the Pythagorean Theorem and the given values of the other two sides (see 1996 Grade 8 TAAS #15). By definition, this skill requires the use of arithmetic operations and is coded by itself. It should be noted that the Pythagorean Theorem is supplied on the formula chart for the Grade 8 TAAS exam and is not provided on the one skill 33 question on NAEP.

***Not used on Grade 4 TAAS exams.*

Skill 33 – Angles and Triangles

Questions coded 34 require knowledge of triangles and their properties as well as general knowledge about angles. These questions may require students to (1) differentiate and identify right, acute and obtuse angles, (2) understand and apply the definition of complementary and supplementary angles, (3) recognize that the three angles of a triangle sum to 180 degrees and/or

use this knowledge to solve for the measure of one angle in a triangle, or (4) differentiate and identify different types of triangles such as right, isosceles or equilateral. Code 34 will be combined with other skill codes, when necessary, and if the question requires the calculation of the value of an angle, the appropriate arithmetic code will be used. Typical questions include identification of the type of angle pictured (see NAEP questions that may ask how many right angles are there in the given picture) and calculation of the value of an adjacent angle; for example #14 on the Grade 8 TAAS in 1996 reads “A very thin book on Chelsea’s bookshelf was leaning against a bookend. It formed a 60° angle with the shelf, as shown in the diagram. Approximately what was the measure of the angle between the shelf and the other side of the book?”.

***Not used on Grade 4 TAAS exams.*

Skill 34 – Venn Diagrams

Questions coded 35 involve Venn Diagrams. To answer these questions correctly, students must be able to understand and apply the information displayed in a Venn Diagram and use it to identify a reasonable conclusion or the number of observations that fit the given criteria. For example, item #39 on the 1996 Grade 8 TAAS exam reads “The Venn diagram shows the relationship of 4 sets of students at Tony’s high school. Each set and the intersection shown has at least 1 member in it. Which is a valid conclusion concerning these sets?”. Skill 35 is only combined with 12 – logical reasoning, where appropriate.

***Not used on Grade 4 TAAS or NAEP*

Skill 35 – Combinations

Questions coded 36 require students to understand and apply combinations. In one question, for example, the following is described “A pizza restaurant offers 5 different meat toppings, 4 different vegetable toppings, and two types of crust. If Lisa chooses 1 meat topping, 1 vegetable topping and 1 crust, how many different combinations can she order?”. Skill 36 is not combined with other skill codes.

***Not used on Grade 4 TAAS or NAEP.*

Skill 36 – Reading a Measurement

These are questions that require students to read a measurement. This may require use of a ruler (NAEP only) or other scale of measure pictured in the question. For example, question #2 on the 1996 Grade 8 TAAS reads “The scale pictured below weighs items in pounds. To the nearest pound, what is the weight of the gold bar shown?”. Questions coded 50 also assume knowledge that to balance a scale the weight must be equal on both sides. This is a code that will be combined with other skill code as needed (for example, a typical question on NAEP may require students to take two measurements and find the difference between them).

Skill 37 – Understanding Units of Measure

Questions coded 51 test students' understanding of measurement and units of measure. Typically, these items will ask students to identify the reasonable weight, size or capacity of a specified object (for example #12 on the 2000 Grade 4 TAAS “Which object is most likely to have a capacity of 30 gallons?”) or to select the appropriate instrument or unit of measure for weight, volume, distance, etc. (for example, a NAEP question may ask what the appropriate unit of measure would be for the length of a small object). Skill code 51 is used by itself.

Skill 38 – Order of Operations

Questions coded 61 require students to understand the use of parentheses to indicate the order in which to perform operations. 61 will be combined with all other skill codes needed to solve the problem correctly.

***Not Used on Grade 4 and 8 TAAS exams..*

Skill 39 – Writing Numerals

Questions coded 62 require knowledge of how numerals are written in words (see 1996 Grade 4 TAAS exam item # 11). Typically, these items will provide a written number and ask students to choose the corresponding numeral from the choices given, or given a numeral, students must select the answer choice with the numeral correctly written in words. This skill code is only combined with a fraction or decimal code, when necessary, to indicate that the numbers involved contain decimals and/or fractions.

***Not used on Grade 8 TAAS exams.*

Skill 40 – Place Value

Questions coded 63 require knowledge of place value. Typically, questions may ask students to identify the place value of a certain digit in a given number or ask students to compare digits across different places (see 1996 Grade 4 TAAS #9 which asks which number has the largest digit in the tens place). Some NAEP questions test students knowledge of place value in a pictorial context. For example, a question may depict sets of shapes each representing a different place value and ask what number is represented. Code 63 will be combined with other skill codes as needed, for example a NAEP question may ask by how much a number will increase if the digit in the tens place is replaced with a different digit .

*** Nor used on Grade 8 TAAS exams.*

Skill 41 – Odd/Even

Questions coded 16 require knowledge of even and odd numbers. To answer these questions correctly, students must identify a group of odd or even numbers from a set of numbers that contains both. For example, #7 on the 1996 Grade 4 TAAS exam reads “Which of the following is a set of odd numbers?” with a correct response of “19, 21, 23, 25”. Code 64 is often used alone as in the example just noted. However, it can be combined with other skill codes, where appropriate. For example, a question on a NAEP exam may provide students with a statement

about x being a number greater than 2 and ask them to identify the expression that represents the next even number.

*** Nor used on Grade 8 TAAS exams.*

Skill 42 – Time

Questions coded 71 require knowledge of time, specifically of measuring time in hours and minutes. Typical questions may give a beginning time and ask what time it is x minutes later (see 1996 Grade 4 TAAS #10) or give the length of time a task takes in hours and ask how many minutes the task takes. Code 71 assumes the ability to add and subtract hours and minutes and is not combined with arithmetic operation codes. It may also test the ability to discuss fractions of an hour as quarters and halves.

*** Nor used on Grade 8 TAAS exams.*

Skill 43 – Shapes and Figures

Questions coded 80 require knowledge of shapes, figures and their properties. Such items may ask students to (1) identify the name of a depicted shape (see 1996 Grade 4 TAAS #4), (2) identify the number of sides/faces for a particular shape (see 1997 Grade 4 TAAS #19). It should be noted that these questions may require students to know and understand terms that describe categories of shapes or figures (e.g., parallelogram, polygon, quadrilateral, hexagon, etc.). Questions coded 80 may be combined with other skill codes as necessary and will most often appear with other geometry or angle related codes (34, 81, etc.).

Skill 44 – Geometry

Questions coded with skill 81 require an understanding and application of certain terms used in geometry. More specifically, these items test student understanding of the following terms: congruence, symmetry, rotation, translation, reflection, line segment, plane and ray. A typical question in this skill category will show three pictures and ask students which depicts the given figure being rotated over an axis or for which figure is the x -axis a line of symmetry. Code 81 can be combined with other codes, as needed, and it should be noted that questions coded 81 often are coded with skill 20 (Cartesian Graphs) on TAAS.

Appendix B – Format Categories

Definitions used to categorize the skills assessed in items on both the Texas Assessment of Academic Skills (TAAS) and the National Assessment of Educational Progress (NAEP) mathematics exams in grades 4 and 8 from 1996 to 2000. Items can have multiple formats, in which case they will have more than one format code.

Code	Definition
W	Question is in the form of a word problem.
C	Question is in the form of a calculation problem – that is, no words.
V	Question is in the form of a vertical calculation problem. This code only applies to arithmetic calculation problems.
H	Question is in the form of a horizontal calculation problem. This code only applies to arithmetic calculation problems.
E	Problem contains extraneous information.
P	Problem includes a picture aid. This code applies, by definition, to all problems involving graphs or charts (skills 20-24 and 35).
D	The relevant definition is provided in the problem.
S	The question requires student to merely identify a strategy for solving the problem, rather than actually solving the problem.
L	Problem involves a number line, which is provided in the question.
G	Problem requires students to use information given in the problem.
	Problem requires students to know the definition of the term “fact family.” This definition only appears in several questions in later years of the 4 th grade TAAS exam.
N	The question is asked in the negative.
M	Problem involves money.
O	Correct response not provided (that is, omitted).
5	Five possible answer choices instead of four.
R	Ruler needed (NAEP only)
A	Aids or manipulatives needed (NAEP only)
T	Protractor needed (NAEP only)
U	Calculator may be used (NAEP only as directed)

Table 1 – OLS Estimates of Average Student Performance in the State

	Math				
	All Years	1992- 2000	2000- 2003	Grades 3-5	Grades 6-8
Annual achievement trend	.048** (.013)	.045** (.014)	.065 (.038)	.061** (.014)	.033** (.013)
Annual achievement trend * State exam	.040** (.018)	.047** (.011)	-0.001 (.043)	.017 (.017)	.065** (.017)
State*grade level fixed effects	Yes	Yes	Yes	Yes	Yes
Cubic in the number of times the test had been administered	Yes	Yes	Yes	Yes	Yes
Number of observations	137	105	57	66	71
R-squared	.846	.836	.766	.772	.913

	Reading				
	All Years	1992- 2000	2000- 2003	Grades 3-5	Grades 6-8
Annual achievement trend	.015 (.009)	.025 (.015)	-.067 (.091)	.022** (.011)	.019** (.006)
Annual achievement trend * State exam	.049** (.021)	.044** (.020)	.115 (.103)	.042 (.029)	.041** (.016)
State*grade level fixed effects	Yes	Yes	Yes	Yes	Yes
Cubic in the number of times the test had been administered	Yes	Yes	Yes	Yes	Yes
Number of observations	116	78	52	73	43
R-Squared	.756	.779	.677	.851	.700

Notes: Standard errors are calculated using a block bootstrap with 1000 replications with state as the blocking variable. * = significance at the 10 percent level, ** = significance at the 5 percent level.

Table 2 - Summary Statistics for TAAS-NAEP Analysis

Variable	TAAS				NAEP			
	Grade 4		Grade 8		Grade 4		Grade 8	
	1996	2000	1996	2000	1996	2000	1996	2000
Percent of multiple-choice items answered correctly	0.80	0.85	0.74	0.80	0.57	0.60	0.56	0.56
Black	0.14	0.15	0.13	0.13	0.14	0.15	0.12	0.13
Hispanic	0.31	0.37	0.33	0.36	0.34	0.36	0.37	0.38
Asian	0.01	0.02	0.02	0.03	0.02	0.04	0.03	0.04
Male	0.50	0.50	0.50	0.50	0.51	0.47	0.47	0.51
Free-reduced lunch	0.47	0.47	0.38	0.38	0.45	0.47	0.39	0.44
Limited English Proficient	0.05	0.11	0.05	0.06	0.10	0.06	0.04	0.05
Special education	0.09	0.07	0.08	0.06	0.06	0.05	0.05	0.07

Notes: NAEP statistics account for sample weights and complex sampling design.

Table 3 – The Fraction of Items that Assess Mastery of Specific Skills

Skill	Fraction of items		Skill	Fraction of items	
	NAEP	TAAS		NAEP	TAAS
<i>Arithmetic</i>			<i>Geometry</i>		
Addition	0.15	0.12	Perimeter	0.01	0.02
Subtraction	0.09	0.14	Volume	0.01	0.01
Multiplication	0.13	0.19	Pythagorean Theorem	0.01	0.01
Division	0.08	0.13	Angles	0.05	0.01
Decimals	0.10	0.17	Shapes & figures	0.01	0.03
Fractions	0.07	0.03	Area	0.03	0.01
Percents	0.03	0.03	Spatial Reasoning	0.05	0.01
Rate/Time/Distance	0.00	0.03	<i>Graphs</i>		
Ratios/proportions	0.01	0.04	Cartesian Graphs	0.02	0.01
Order of operations	0.01	0.00	Line graphs	0.01	0.01
Writing numerals	0.01	0.01	Pie charts	0.04	0.02
Place value	0.02	0.00	Bar Graphs	0.00	0.04
Odd/even	0.02	0.00	Other graphs/tables	0.02	0.05
Negative Numbers	0.01	0.01	Venn Diagrams	0.00	0.00
Greater than/Less than	0.02	0.03			
Rounding	0.02	0.01			
<i>Statistics</i>			<i>Algebra</i>		
Probability	0.04	0.02	Writing algebraic expressions	0.01	0.04
Statistics	0.02	0.01	Solving algebraic expressions	0.08	0.00
Combinations	0.03	0.01			
<i>Measurement</i>			<i>Exponents</i>		
Reading a measurement	0.06	0.01	Exponents	0.01	0.01
Understanding units of measurement	0.00	0.00	Exponential notation	0.01	0.00
Measurement Conversion	0.01	0.01			
<i>Other</i>					
Estimation	0.01	0.11			
Logical Reasoning	0.03	0.02			
Pattern Recognition	0.04	0.01			
Time	0.09	0.03			

Note: Only includes multiple-choice items. Data includes items from grades 4 and 8 in years 1996 and 2000. NAEP statistics account for sample weights and complex sampling design.

Table 4 – The Fraction of Items that Assess Mastery of Specific Formats

	Fraction of Items	
	NAEP	TAAS
<i>Format</i>		
Word problem	0.46	0.66
Calculation problem – no words	0.04	0.02
Vertical calculation	0.01	0.00
Horizontal calculation	0.03	0.02
Extraneous information	0.01	0.04
Picture aide	0.34	0.20
Definition provided	0.00	0.00
Identify solution strategy; do not solve	0.06	0.11
Includes number line	0.01	0.01
Use given information	0.00	0.04
Asked in the negative.	0.03	0.02
Problem involves money.	0.02	0.02
Correct response not provided (that is, omitted).	0.00	0.02
Five possible choices are provided (instead of four)	0.00	0.36
<i>Use of Aides</i>		
Calculators	0.29	0.00
Manipulatives	0.06	0.00
Rulers/protractors	0.05	0.00

Note: Only includes multiple-choice items. Data includes items from grades 4 and 8 in years 1996 and 2000. NAEP statistics account for sample weights and complex sampling design.

Table 5 – OLS Estimates of Item Performance

	Dependent Variable = Correctly answered item					
	Full Sample				4 th Grade	8 th Grade
	(1)	(2)	(3)	(4)	(5)	(6)
Year 2000	0.044** (0.003)	0.012 (0.008)	0.016** (0.006)	-0.036** (0.006)	-0.019** (0.008)	-0.050** (0.009)
Grade 8	-0.048** (0.003)	-0.048** (0.003)	-0.057** (0.002)	-0.057** (0.002)	--	--
TAAS	0.225** (0.004)	0.201** (0.006)	0.205** (0.004)	0.206** (0.004)	0.234** (0.006)	0.184** (0.006)
Year 2000*TAAS		0.045** (0.008)	0.044** (0.006)	0.043** (0.006)	0.026** (0.008)	0.056** (0.009)
Item position – 2 nd quartile of items			-0.077** (0.001)	-0.103** (0.002)	-0.102** (0.002)	-0.104** (0.003)
Item position – 3 rd quartile of items			-0.139** (0.002)	-0.173** (0.002)	-0.130** (0.002)	-0.206** (0.003)
Item position – 4 th quartile of items			-0.066** (0.001)	-0.112** (0.002)	-0.109** (0.002)	-0.115** (0.002)
2 nd quartile * Year 2000				0.051** (0.002)	0.081** (0.003)	0.026** (0.004)
3 rd quartile * Year 2000				0.065** (0.003)	0.053** (0.003)	0.075** (0.004)
4 th quartile * Year 2000				0.092** (0.003)	0.059** (0.003)	0.117** (0.004)
Black			-0.118** (0.003)	-0.118** (0.003)	-0.104** (0.004)	-0.129** (0.004)
Hispanic			-0.057** (0.002)	-0.057** (0.002)	-0.044** (0.003)	-0.067** (0.003)
Asian			0.053** (0.006)	0.053** (0.006)	0.042** (0.009)	0.057** (0.007)
Male			0.009** (0.002)	0.009** (0.002)	0.009** (0.002)	0.009** (0.002)
Eligible for free or reduced-price lunch			-0.055** (0.002)	-0.055** (0.002)	-0.060** (0.003)	-0.052** (0.003)
Limited English proficiency			-0.095** (0.004)	-0.095** (0.004)	-0.074** (0.006)	-0.118** (0.006)
Special education			-0.147** (0.004)	-0.147** (0.004)	-0.132** (0.005)	-0.161** (0.006)
Number of observations	2,939,063	2,939,063	2,939,063	2,939,063	1,302,177	1,636,886
R-squared	0.058	0.058	0.101	0.103	0.104	0.101

Note: The sample includes students in grades 4 and 8 in years 1996 and 2000 who took the TAAS or the NAEP. Only multiple-choice items are included.

Table 6 – OLS Estimates of Item Performance, Alternative Specifications

Specification	$\Delta Gap =$ $(TAAS_{00} - TAAS_{96}) -$ $(NAEP_{00} - NAEP_{96})$	Controls for student demographic s and item position	Controls for 14 skill categories	Controls for 8 format categories	Controls for 14 skill categories and 8 format categories	Controls for 145 categories that capture skills and formats together
	(1)	(2)	(3)	(4)	(5)	(6)
<u>4th Grade (n=1,302,177)</u>						
Year 2000*TAAS	0.021* (0.011)	0.026** (0.008)	0.042** (0.009)	0.006 (0.009)	0.028** (0.009)	0.006 (0.010)
R-squared	0.067	0.104	0.117	0.111	0.124	0.153
<u>8th Grade (n=1,636,886)</u>						
Year 2000*TAAS	0.065** (0.012)	0.056** (0.009)	0.060** (0.009)	0.055** (0.010)	0.062** (0.010)	0.024** (0.009)
R-squared	0.048	0.101	0.114	0.109	0.119	0.156

Note: The sample includes students in grades 4 and 8 in years 1996 and 2000 who took the TAAS or the NAEP. Only multiple-choice items are included.

Table 7 – OLS Estimates of Item Performance, by Race and Gender

Model	White (1)	Black (2)	Hispanic (3)	Male (4)	Female (5)
<u>4th Grade</u>					
Controls for demographics and item position	0.020* (0.011)	0.042** (0.016)	0.030** (0.013)	0.038** (0.012)	0.018* (0.010)
Controls for 14 skill categories and 8 format categories	0.017 (0.012)	0.038** (0.018)	0.025* (0.013)	0.028** (0.012)	0.020* (0.011)
Controls for 145 categories that capture skills and formats together	-0.017 (0.012)	0.031 (0.021)	0.010 (0.015)	0.014 (0.013)	-0.012 (0.011)
<u>8th Grade</u>					
Controls for demographics and item position	0.044** (0.010)	0.083** (0.020)	0.064** (0.013)	0.072** (0.011)	0.044** (0.010)
Controls for 14 skill categories and 8 format categories	0.090** (0.011)	0.180** (0.024)	0.135** (0.014)	0.136** (0.012)	0.103** (0.012)
Controls for 145 categories that capture skills and formats together	0.044** (0.011)	0.075** (0.022)	0.052** (0.015)	0.065** (0.012)	0.043** (0.011)

Note: The sample includes students in grades 4 and 8 in years 1996 and 2000 who took the TAAS or the NAEP. Only multiple-choice items are included.

Figure 1: Student Achievement Trends on NAEP and State Assessments in Texas

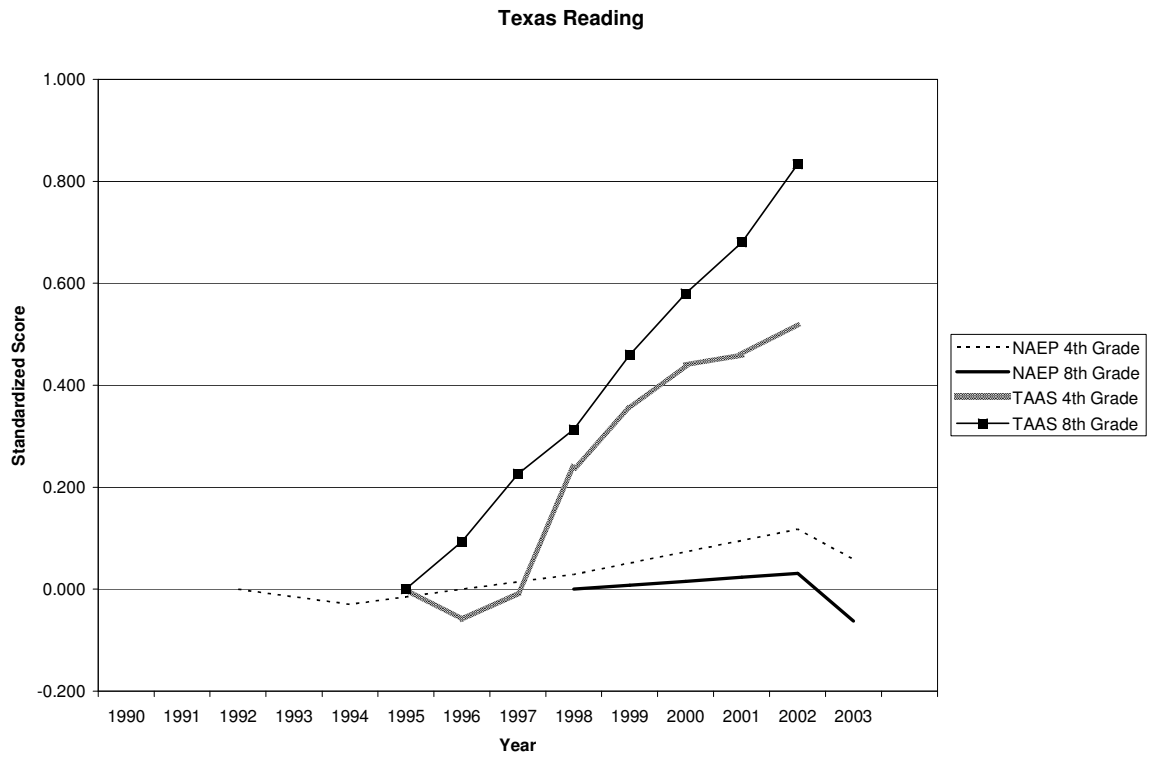
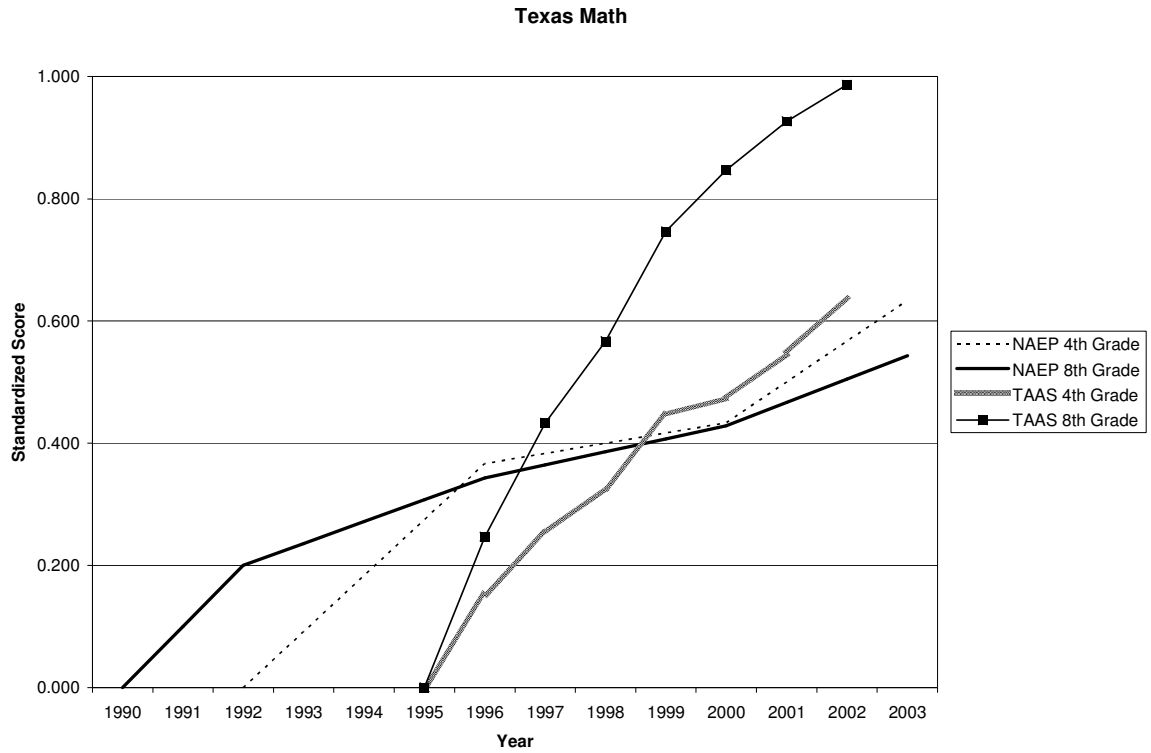


Figure 2: Student Achievement Trends on NAEP and State Assessments in North Carolina

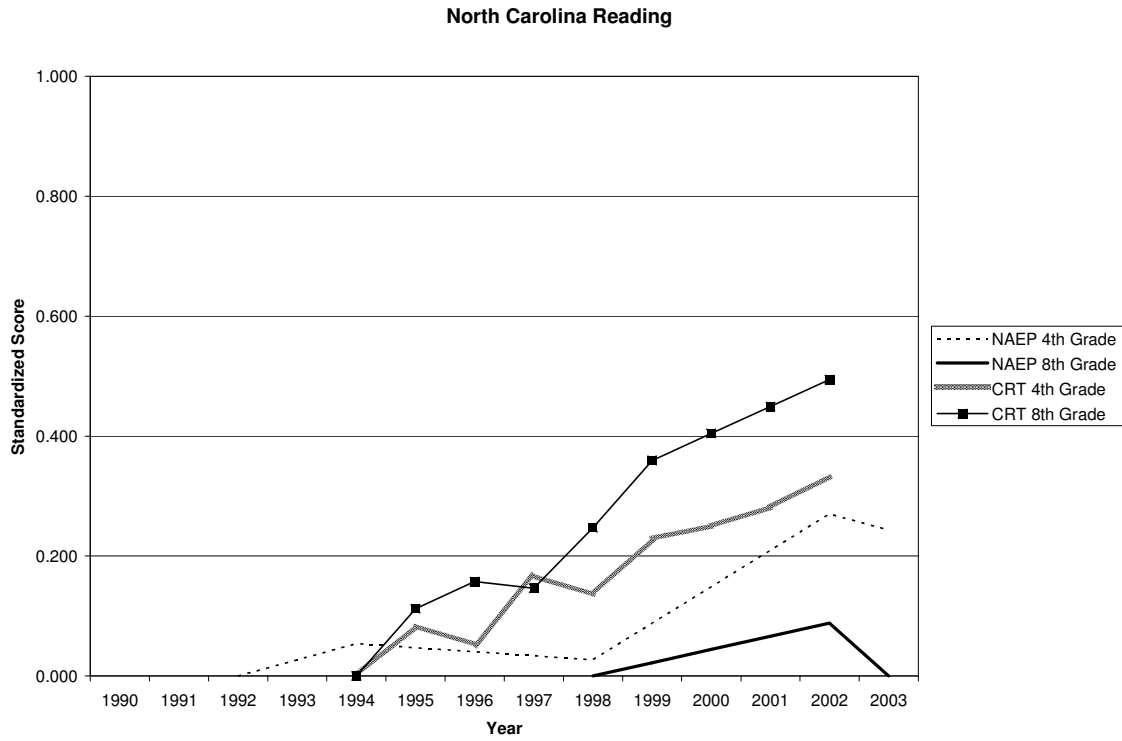
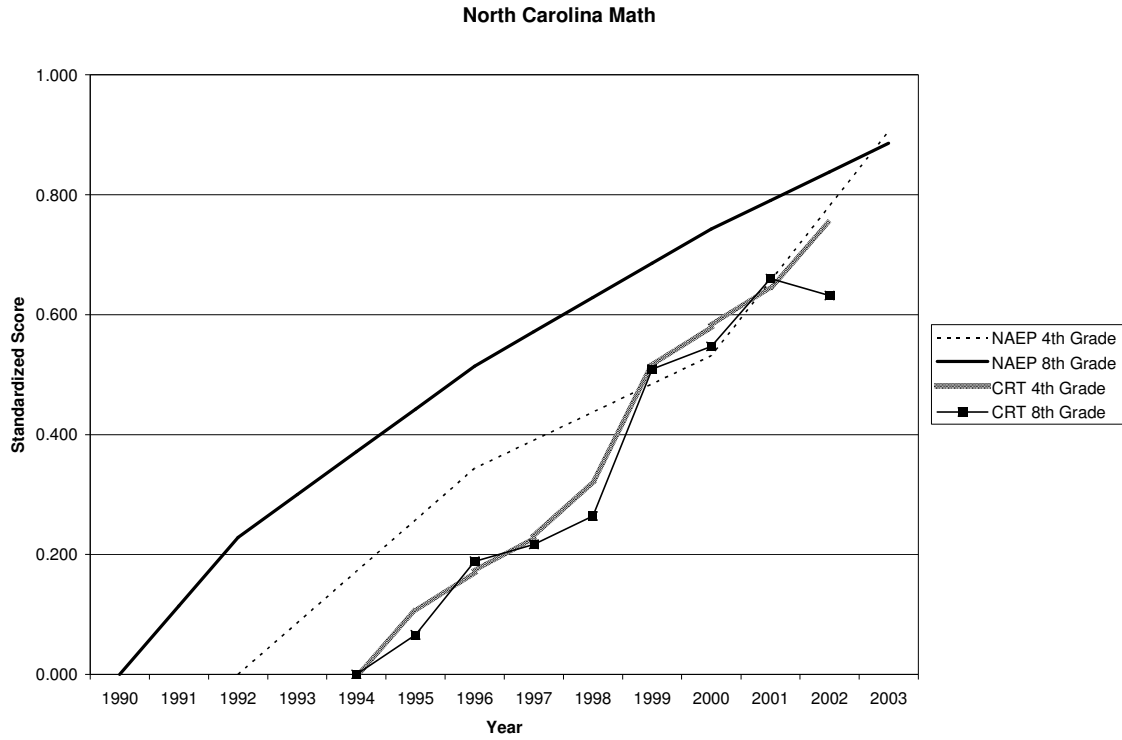


Figure 3: Student Achievement Trends on NAEP and State Assessments in Connecticut

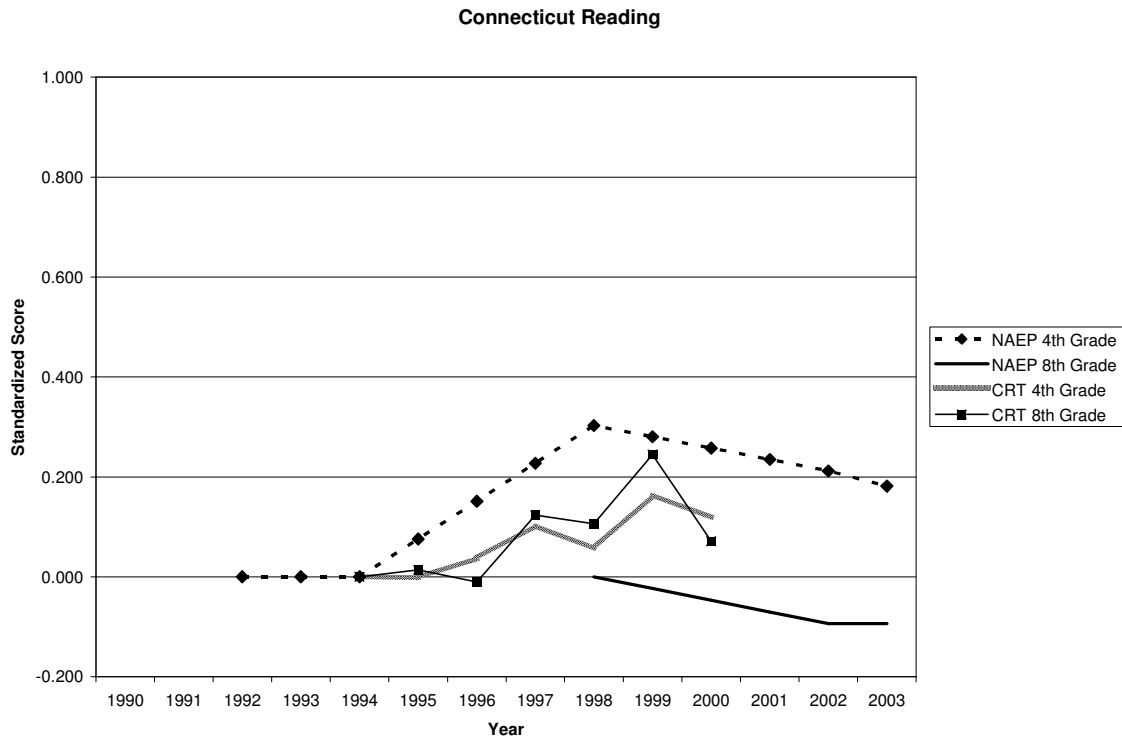
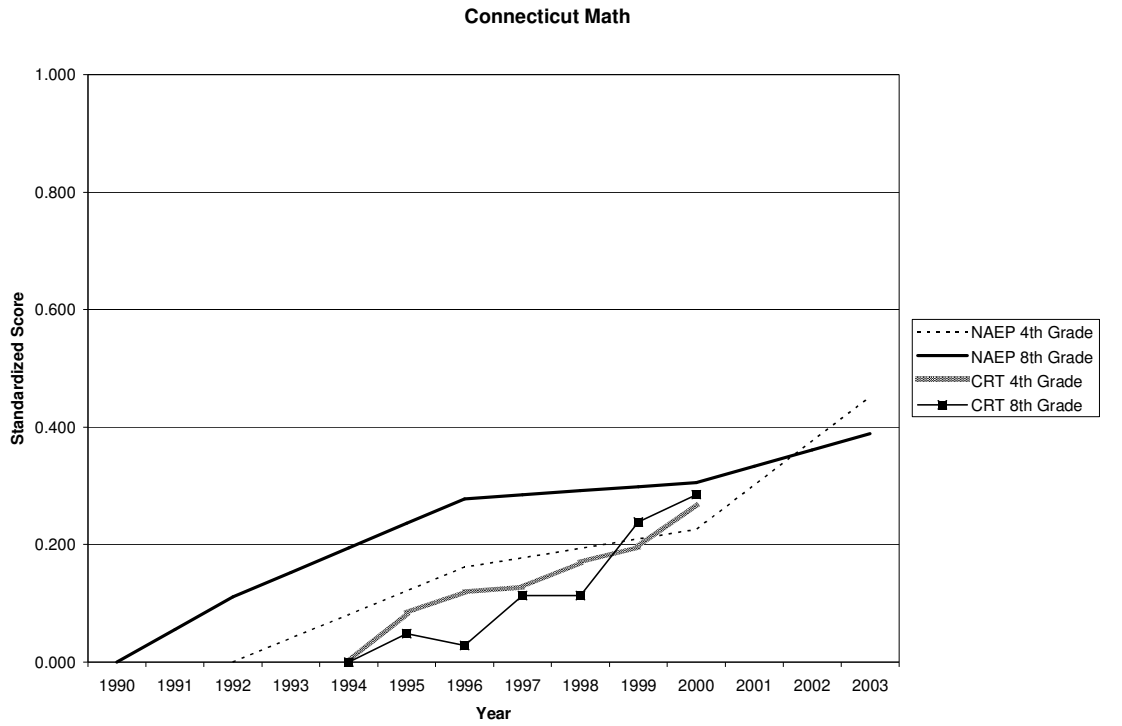


Figure 4: Student Achievement Trends on NAEP and State Assessments in Arkansas

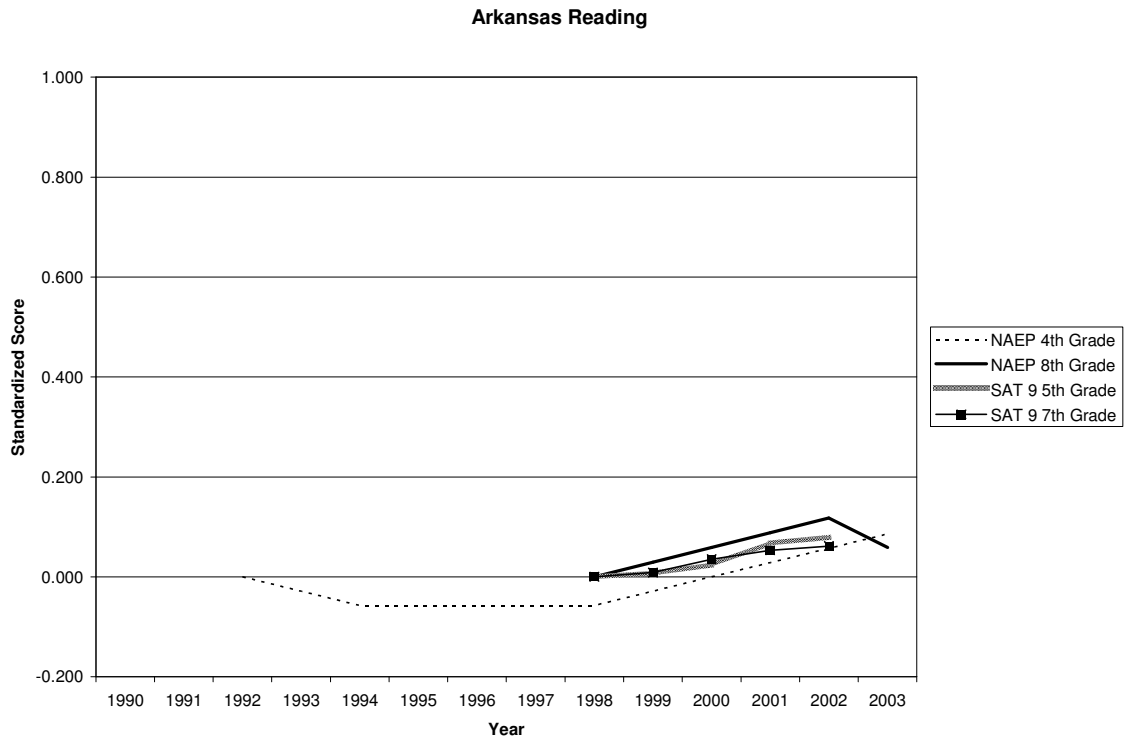
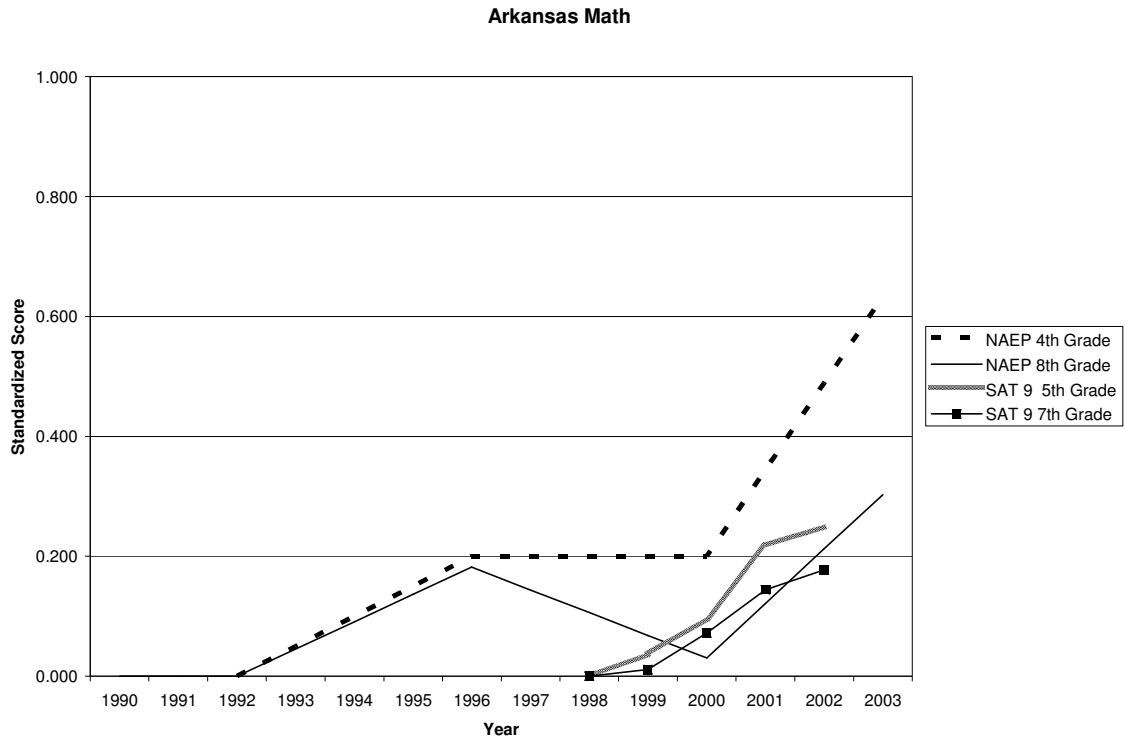


Figure 5: Texas 4th Grade Math Performance

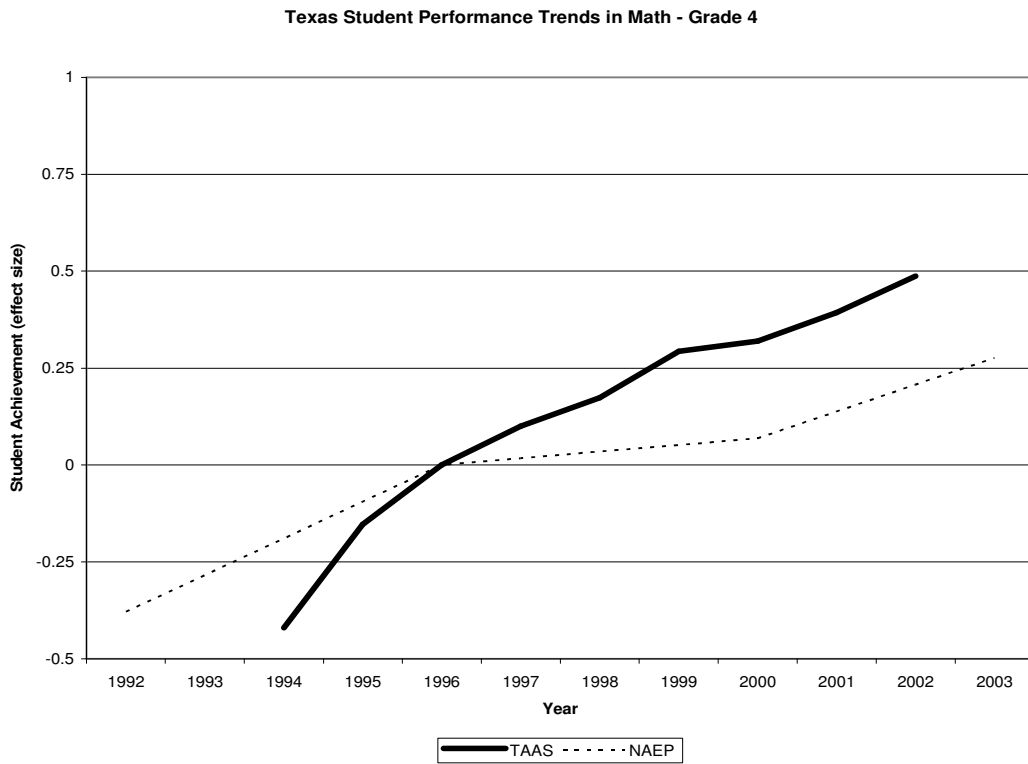


Figure 6: Texas 8th Grade Math Performance

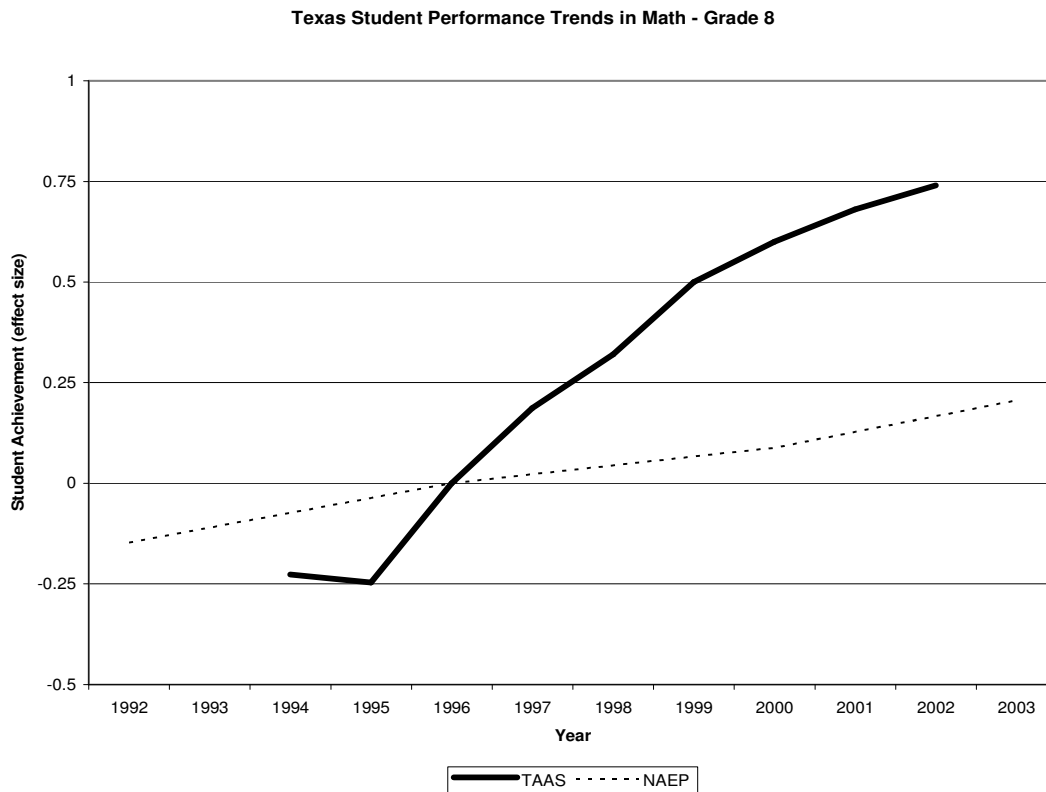


Figure 7: Texas 4th Grade Math Performance on Multiple-Choice Items Only

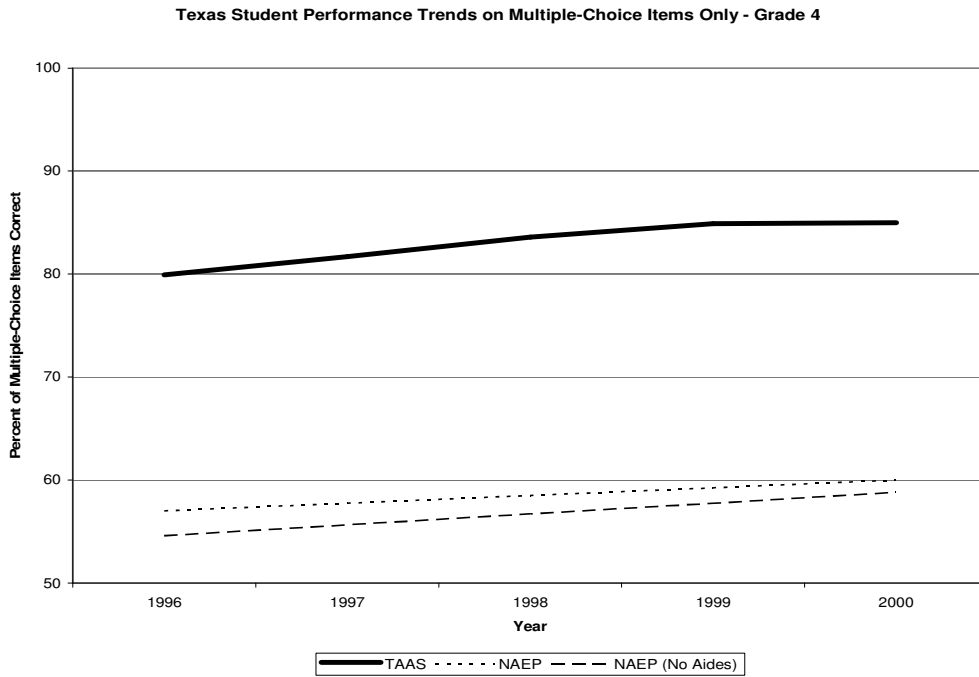


Figure 8: Texas 8th Grade Math Performance on Multiple-Choice Items Only

