DOCUMENT RESUME

ED 476 420 TM 034 922

AUTHOR	Jacob, Brian A.
TITLE	Test-Based Accountability and Student Achievement Gains: Theory and Evidence.
REPORT NO	PEPG/02-06
PUB DATE	2002-06-00
NOTE	23p.; Paper presented at a conference at the John F. Kennedy School of Government, Harvard University (Cambridge, MA, June 10-11, 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Accountability; *Achievement Gains; Educational Change; Elementary Secondary Education; *Grade Inflation; School Districts; *Scores; *Test Use
IDENTIFIERS	*Chicago Public Schools IL

ABSTRACT

This paper examines the issue of test score inflation in the context of test-based accountability. The first section provides some background on the topic, describing what exactly is meant by test score inflation, reviewing the existing evidence for such inflation, and discussing why one should or should not be concerned if scores are inflated. When test scores are inflated, they are no longer good indicators of overall student skills. The second section discusses a number of ways to understand whether test score gains resulting from an accountability policy are meaningful. The third section presents some evidence on the factors driving test score improvements in Chicago following the implementation of high-stakes testing in that district. Data from the school system were obtained for students in grades 3, 6, and 8 from 1993 to 2000. The final section discusses the implications of potential test score inflation for education policy. Carefully examining the nature of gains on a high-stakes examination can reveal much about real performance changes under an accountability program. (Contains 9 tables and 28 references.) (SLD)

Test-Based Accountability and Student Achievement Gains: Theory and Evidence*

Brian A. Jacob
Harvard University and NBER

PEPG/02-06

A paper prepared for the conference
Taking Account of Accountability: Assessing Politics and Policy
John F. Kennedy School of Government
Harvard University
June 10 - 11, 2002

* Preliminary work. Please do not cite or reprint. I would like to thank Robin Tepper and Lars Lefgren for helpful comments and the Chicago Public Schools and Consortium on Chicago School Research for providing the data used in this study. All remaining errors are my own.

Introduction

In January 2002, President Bush signed the "No Child Left Behind" Act of 2001, ushering in a new era of educational accountability. School reforms designed to hold students and teachers accountable for student achievement are already in place throughout the country. Statutes in 252 states explicitly link student promotion to performance on state or district assessments. At the same time, 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 (Quality Counts, 2002).

Accountability advocates claim that such policies will motivate students and teachers to work harder, cause parents to become more involved in their children's education and force school administrators to implement more effective instruction. Pointing to Texas, North Carolina and Chicago, they argue that test-based accountability can substantially improve student learning. Critics of test-based accountability respond that such policies lead to a host of undesirable outcomes, including a narrowing of the curriculum, a shift away from low-stakes subjects and untested skills and an increase in behaviors designed to game the system, such as placing low-ability students in special education where they will either not be tested, or receive special accommodation on the exams.

Perhaps the most serious criticism of high-stakes testing is that it leads to "inflated" test scores that do not truly reflect students' knowledge or skills and therefore cannot be generalized to other outcome measures. This issue received national publicity when the RAND Corporation released a study during the last presidential campaign indicating that Texas students improved much less during the 1990s on the National Assessment of Educational Progress (NAEP) than the state TAAS (Texas Academic Assessment System) exam (Klein et. al. 2000).

This paper examines the issue of test score inflation in the context of test-based accountability. The first section provides some background on the topic, describing what exactly is meant by test score inflation, reviewing the existing evidence for such inflation and discussing why one should or should not be concerned if test scores are inflated. The second section discusses a variety of ways to better understand whether test score gains resulting from an accountability policy are meaningful. The third section presents some evidence on the factors driving test score improvements in Chicago following the introduction of high-stakes testing in that district. The final section discusses the implications of potential test score inflation for education policy.

Conceptual Framework

What is 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 be "smart" or 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 understanding of the subject matter.

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 number of reasons why test score inflation could occur. While cheating is undoubtedly the most egregious cause of score inflation, test preparation is perhaps the most common. When teachers focus instruction on particular topics and skills that are commonly measured on the high-stakes exam, students may make substantial improvements on the exam because of their improvement on these specific items, rather than a general improvement in the larger subject area.

Evidence of Test Score Inflation

There is considerable evidence of test score inflation during the past two decades. In 1987, Cannell (1987) discovered what has become known as the "Lake Wobegon" effect—the fact that a disproportionate number of states and districts report being "above the national norm." This phenomenon was documented in several studies, one of which concluded that teaching to the test played a role in these results (Cannell, 1987, Linn et. al. 1990, Shepard 1990). Linn and Dunbar (1990) found that states have made smaller gains on the National Assessment of Educational Progress (NAEP) than their own achievement exams.

There is less evidence on whether, and to what extent, accountability programs lead to test score inflation. 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. One particularly interesting finding was that scores dropped sharply when a new form of test was introduced and then rose steadily over the next several years as teachers and students became more familiar with the exam.

Koretz and Barron (1998) examined the generalizability of gains on the Kentucky Instructional Results Information System (KIRIS) testing program. They not only examined test score patterns internal to the KIRIS, but also compared performance gains on KIRIS to gains on other assessments over the same time period. They conclude that KIRIS gains were quite large during the initial years of the program but did not generalize to performance on the NAEP or the ACT. 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. (2001) conducted a similar analysis, comparing performance trends of Texas students in the 1990s on both the NAEP and the TAAS.

Should We Care about Test Score Inflation?

In discussing the meaningfulness of test score gains, Koretz (forthcoming) states, "When scores increase, students clearly have improved the mastery of the sample included in the test. This is of no interest, however, unless the improvement justifies the inference that students have attained greater mastery of the domain the test is intended to represent." While this is certainly true in the extreme, it perhaps neglects the importance of actual improvement on specific test items. If children truly improve their ability to add fractions, interpret line graphs or identify the main idea of a written passage, is this of *no* interest?

Most importantly, this statement illustrates the importance of understanding the nature of any test score gain, particularly for the purposes of educational policy. The question is less whether test score are inflated, and thus not generalizable, but more whether they are meaningful.[1] One way to think about whether or not gains are meaningful is to better understand the factors underlying the gains. There are four factors that might produce test score gains under high-stakes testing: improvement in general skills, improvement in test-specific skills, increases in student effort or other testing conditions, or cheating. These underlying causes have different implications for how we would interpret the meaningfulness of the test score gains.

At one extreme, test score gains may be the result of cheating on the part of students or teachers. While cheating may seem unusual, documented cases of such cheating have recently been uncovered in California (May 2000), Massachusetts (Marcus 2000), New York (Loughran and Comiskey 1999), Texas (Kolker 1999), and Great Britain (Hofkins 1995, Tysome 1994). Jacob and Levitt (2002) find that teacher cheating is extremely responsive to incentives. In their study, the introduction of a test-based accountability program increases the prevalence of cheating by roughly 50 percent. If test score gains were driven entirely by cheating, most people would consider the apparent improvements in achievement as completely without merit.

Another potential explanation for achievement gains involves improvements in testing conditions or increases in student effort on the day of the exam (as distinct from effort throughout the school year). Unless the higher effort was indicative of a more serious attitude to performance in general, one would probably view these gains are largely meaningless. (Whether a student *could* theoretically do well is less important for life success than whether that student chooses to do well.)

The third reason for a test score gain is an increase in certain specific skills. Here it is important to distinguish again between meaningful and meaningless gains. By definition, an increase in only certain specific skills in a domain—e.g., the ability to add fractions—will not be completely generalizable to other exams. However, to the extent that the newly learned ability to add fractions is meaningful, it may still be a valid outcome. Whether or not observed test score gains reflect true learning of the specific skill or not is an empirical question that is, unfortunately, often difficult to answer.

On one hand, improving test score gains could be due to improvements on the exact question format as well as type contained in the exam, without any deeper understanding of the underlying questions. This is close to rote memorization. In these cases, we would likely consider the improvements meaningless. A dramatic 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. For subtraction of decimals, the passing rates were 78 and 30 percent. On

[1] Koretz (forthcoming) emphasizes this distinction as well.

the other hand, increases in test-specific knowledge could involve learning how to add fractions or some other specific type of question commonly measured on the high-stakes exam. To the extent that students truly learn these concepts, how much weight one gives them depends on how you value these skills.

Finally, test score gains may be due completely to an improvement in general skills, in which case students would make equal gains across all areas in the domain. One would think these results should be generalizable to other outcome measures.

How to Tell Whether Test Score Gains are Meaningful

Given the danger of test score inflation, how can one determine whether and to what extent the test score gains generated by an accountability program are meaningful? Perhaps the most common strategy is to compare student performance trends across exams, as has been done by Koretz and Barron (1998), Klein et. al. (2001) and Jacob (2002). In fact, NCLB requires states to use NAEP to verify gains on their own accountability exams. The notion here is that if the test score gains on the high-stakes exam are not accompanied by gains on other achievement exams, then the gains may not be meaningful. However, 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 the NAEP influences the black-white achievement gap (Koretz, forthcoming).

While this is a sensible idea in theory, it has several drawbacks. First, this strategy requires the existence of two or more achievement exams covering the same subjects and given to the same grades, administered before and after the introduction of the accountability policy. Unfortunately, when states administer multiple exams, they often explicitly administer them to different grades or in different subjects, in order to minimize the amount of testing experienced by any one group of students. For the same reason, many districts phase out older exams when newer accountability exams are introduced. Second, when multiple exams exist, they are often given under different testing conditions, with considerably greater pressure associated with performance on the higher-stakes exam. In this case, if we see greater gains on one exam, it is difficult to disentangle effort from learning. In the extreme, "real" student learning will not show up on the low-stakes exam because of a decrease in student effort. Perhaps more realistically, there is some interaction between effort and learning, whereby greater effort is needed to make greater learning visible. Finally, even if there are two exams given under roughly comparable testing conditions, it is difficult to quantify the expected or acceptable degree of divergence on the student outcomes. As discussed above, because of inherent differences in the domains and item samples across exams, even a well functioning accountability policy should not lead to equivalent gains on a different exam with a different domain and sample of questions. However, there is the question of how big a difference should be a cause for concern? This clearly depends on the how different the exams are, but is nonetheless difficult to pinpoint.

An alternative strategy is to examine changes in other student outcomes. For example, one might examine college entrance exams such as the ACT or SAT, school attendance, grades, high-school graduation or matriculation to college. There are advantages and disadvantages to using each of these outcomes. College entrance exams suffer from selection issues because they are not mandatory and, moreover, can only be used to examine achievement for certain groups of students—e.g., college bound high school students. Grades are a relatively subjective measure that might change in response to the policy – e.g., a get-tough accountability policy may lead to tougher grading standards. Attendance, graduation and college completion are more objective, but capture a different set of skills, abilities and motivations than achievement exams. These might be interesting and important outcomes to examine in their own right, but will not necessarily help us interpret achievement gains.

A related proposal is to examine how the correlation between exams changes over time. The intuition behind this approach is that if the scores on test X become less meaningful because of test score inflation, they will be less predictive of outcomes on test Y. Under this assumption, a decrease in the correlation casts doubt on the meaningfulness of test score gains. While this strategy sounds plausible, it can be shown that gains (or losses) on only one exam will generally *not* change the correlation between exams at all. Denote the initial correlation between the two exams is $r_{XY} = \frac{Cov(X,Y)}{s_X s_Y}$. Suppose that following the introduction of an accountability policy achievement on one the high-stakes exam is $X' = a + bX$ while achievement on test Y remains the same. Note that this simple linear transformation captures a variety of interesting cases, including a scenario in which all students make equal gains ($b = 0$) as well as a scenario in which lower or higher achieving students make relatively larger gains ($0 \le b < 1$ or $b > 1$, respectively). Because the level shift in achievement, *a*, changes neither $Cov(X,Y)$ nor s_X, it will have no influence on the correlation. As shown below, proportional achievement shifts, *b*, will cancel out, again leading to no change:

$$r'_{XY} = \frac{Cov(X',Y)}{s'_X s_Y} = \frac{Cov(a+bX,Y)}{\sqrt{Var(a+bX)}s_Y} = \frac{bCov(X,Y)}{\sqrt{b^2Var(X)}s_Y} = \frac{bCov(X,Y)}{bs_X s_Y} = r_{XY}.$$

Moreover, it can be shown that the changes in measurement error likely to be induced by the introduction of an accountability policy will influence the correlation, independent of any changes in actual achievement on either test. For example, if the accountability policy decreases the prevalence of guessing, it will generally reduce the measurement error on the exam, which will increase the correlation across exams. On the other hand, cheating or other factors could increase measurement error, leading to a decrease in the correlation that does not stem from differential achievement gains.

A fourth strategy avoids the comparisons across different exams or achievement measures, and attempts to look more closely at changes on the accountability exam to provide more information on the changes in learning. Most exams have a collection of items measuring different skills and concepts. By comparing improvement across item type, one can gain some information. For example, if one found that the aggregate achievement gain on a particular exam was comprised of an extremely large gain on computation items, a moderate gain on data interpretation/graphical analysis items and a moderate decline on word problems. Alternatively, one might find a zero aggregate effect on reading comprehension exam is composed of substantial increases in the ability to identify the main idea of a passage and infer an traits, feelings and motives, but a decrease in the ability to recognize structure or style or interpret non-literal language. By examining item strings, it is possible to determine whether there have been changes in guessing or mistakes.

Finally, one might look at changes in potential inputs. If test score gains are meaningful, one should be able to identify changes in inputs that are plausibly related to learning (e.g., change in class size, shift in curriculum, improvement in classroom instruction, etc.). If one found no change in any factor correlated with achievement, one might suspect that the observed achievement gains were driven by factors such as student test-day effort or cheating. The weakness of this approach is that input changes may be extremely difficult to observe. For example, the accountability policy may operate largely through increasing the focus level of students and teachers in the classroom, improving what educators refer to as "time on task." However, to confidently identify changes in time on task, one would likely need systematic, district-level classroom observation data from before as well as after the introduction of the policy.

Evidence from High-Stakes Testing in Chicago

The Chicago Public Schools (ChiPS) was one of the first large, urban school districts to implement high-stakes testing. In 1996-97, the ChiPS introduced a comprehensive accountability program that incorporated incentives for both students and teachers. Beginning in 1996, Chicago schools in which fewer than 15 percent of students met national norms in reading were placed on probation. If student performance did not improve in these schools, teachers and administrators were subject to reassignment or dismissal. At the same time, the ChiPS took steps to end "social promotion," the practice of passing students to the next grade regardless of their academic ability. Students in third, sixth and eighth grades were required to meet minimum standards in reading and mathematics in order to advance to the next grade.

Using a panel of student-level, administrative data, Jacob (2002) found that math and reading achievement on the high-stakes exam (the Iowa Test of Basic Skills or ITBS) increased sharply following the introduction of the accountability policy, in comparison to both prior achievement trends in the district and to changes experienced by other large, urban districts in the mid-west. Student performance on a similar, state-administered achievement exam improved throughout the decade, but did not show a significant deviation from pre-existing trends following the introduction of the accountability policy, calling into question the large ITBS gains. Jacob (2002) presents evidence that the ITBS gains were driven largely by increases in test-specific skills and student effort.[2]

In order to further judge whether the test score gains generated by the accountability policy in Chicago were meaningful, this section uses detailed, item-level test score data to further explore the nature of the ITBS gains in Chicago. More specifically, I will first examine whether there were significant changes in the prevalence of guessing and/or leaving items blank and, if so, whether these changes alone could explain the magnitude of the gains. I next examine which topics and skills showed the largest and smallest gains. Finally, I examine whether there is a relationship between the item difficulty and the gains.

Methodology

The data for this analysis is drawn from student administrative records from the ChiPS, which include not only student test scores and demographic information, but also the actual item

[2] Jacob (2002) also found that teachers responded strategically to the incentives along a variety of dimensions—by increasing special education placements, preemptively retaining students and substituting away from low-stakes subjects like science and social studies.

strings for each student. Unique student identification numbers allow one to follow students over time. The sample includes students in grades three, six and eight from 1993 to 2000. I only consider those students who were tested and whose scores were included for official reporting purposes, which excludes a proportion of bilingual and special education students. The sample is further limited to students who were in the particular grade for the first time, thereby excluding all retained students. The descriptive statistics below generally focus on the 1994 and 1998 cohorts because these groups were well before and after the introduction of the policy and the same form of the ITBS exam was administered in both years. For simplicity, the descriptive and regression estimates often focus solely on the 1998 cohort of eighth grade students. Unless otherwise noted, the results are the same for all other grades and cohorts.

Test Completion, Guessing and Gains

One way in which test-based accountability might influence student achievement is by increasing student effort on the day of the exam. Since there is no penalty for guessing on the ITBS (total score is determined solely by the number correct), the simplest way for a student to increase his or her expected score is to make sure that no items are left blank. Prior to the introduction of the accountability policy in Chicago, a surprisingly high proportion of students left one or more items of he ITBS exam blank. For example, Table 1 shows that in 1994 only 58 and 77 percent of eighth grade students completed the entire math and reading exams respectively. (The higher completion rates in reading are likely due to the fact that it is considerably shorter than the math exam, which consists of three separate subsections.) Most students did not leave many items blank. On average, students only answered 97 percent of the questions on both exams.

As one would expect (and even hope), test completion rates increased sharply under the high-stakes testing regime. The number of eighth graders who completed the entire math exam (i.e., left no blank items) increased from 58 percent in 1994 to nearly 63 percent in 1998, an increase of 5.5 percentage points. There is also evidence that increased guessing played a large role in the improving test completion rates. The percent of students with no blanks at the end of the exam (i.e., final blanks) increased by approximately 9.5 and 6.0 percentage points on math and reading respectively. More direct evidence of guessing is a string of identical responses for the final questions on the exam (e.g., AAA or CCC). Because students may guess in a variety of other ways, the prevalence of identical final strings will generally understate the true level of guessing. It is still instructive to look at how this measure of guessing changed over time. Table 1 shows that the instances of guessing on math increased from 17.1 to 23.4 percent between 1994 and 1998, an increase of roughly 37 percent. Guessing in reading increased by 93 percent.

The bottom panels in Table 1 present similar statistics according to prior student achievement. Prior achievement here is measured as the average fifth grade math and reading score on the ITBS exam. Students who scored in the bottom quartile on a national distribution are considered low ability. Students who scored in the second quartile (26th to 50th percentiles) are classified as moderate ability. Students who scored above the national average are denoted high ability. We see that all of the patterns are replicated across the prior achievement groups. Not surprisingly, the greatest impact was for low-achieving students, largely because nearly all higher-achieving students had always finished the exam.

What percentage of the observed achievement gains in Chicago can be explained solely on the basis of the increase in guessing? If we believe that the increased test scores were due solely to guessing, we might expect the percent of questions answered to increase, but the percent of questions answered *correctly* (as a percent of all *answered* questions) to remain constant or perhaps even decline. Table 1 shows that the percent of questions answered has increased, but that the percent answered correctly has also gone up, suggesting that the higher completion rates were not due entirely to guessing.

Table 2 shows how much student achievement would have increased if students had randomly filled in answers for all of the questions that they would have left blank under a low-stakes regime. I focus on the lowest-achieving group of students in the CPS—those who scored below the bottom quartile—since these students showed the largest increases in test completion under high-stakes testing. The first column reports OLS estimates of the additional number of questions students completed on the ITBS exam under high-stakes testing. If these increases were due solely to guessing, then we would expect students to get 25 percent of the items correct (there are four choices for each question), yielding the number of additional correct responses shown in column two. The third column shows the achievement gain associated with an additional correct response on the ITBS, based on a simple OLS regression of achievement score on the number of correct items.[3] The fourth column shows the achievement gain associated with higher guessing rates on the ITBS under high-stakes testing. Column 5 shows the observed ITBS gain over this period in grade equivalents. The final column shows the maximum fraction of the gain that could be explained by guessing. For all subjects and grades, we see that even this quite generous estimate suggests that only 5 to 15 percent of the gains could be due to guessing.

Table 3 presents regression estimates that are largely comparable to the descriptive statistics shown in Table 1. In addition for controlling for observable demographic characteristics and time trends[4], the estimates in this table are broken out by separate math subsection. Interestingly, it appears that there was the least change in the section on problem solving, perhaps because this section had the fewest questions.

Achievement Gains by Item Type and Difficulty

Having examined the patterns in test completion and guessing, I now examine how student achievement gains differed by item type and item difficulty. This will shed light on how generalizable the results are likely to be. Table 4 presents OLS estimates of the relationship between high-stakes testing and student achievement on the ITBS exam. In the first two rows, we see that students scored roughly 0.20 standard deviation higher following the introduction of high-stakes testing relative to peers prior to high-stakes testing. Rows 3 to 5 show the effects for the three different math subtests. While student performance improved in all subsections, the largest gains were on the section containing math computation questions. The smallest gains came on the subsection on problem-solving and data interpretation. One might ask whether the larger gains in computation could be driven by the fact that students were doing very poorly on this prior to the accountability policy. This turns out not to be the case. The bottom panel presents the district-level averages (in grade equivalents) for 1994 in order to provide a sense of the baseline from which students started. Chicago students performed somewhat better in math than reading. Within mathematics, students had higher achievement on the computation subsection, relative to the number concepts and problem-solving sections.

[3] While the relationship is not perfectly linear, these regression explain over 90 percent of the variance in achievement scores. Using estimated gains based on a more complex relationship between the number correct and achievement does not change the results significantly.

[4] For a complete list of variables included in the regression, see the notes at the bottom of the table.

Table 5 shows the trends in achievement by item type for math. The cells in the first three columns present the proportion of students correctly answering the particular type of question. For example, we can see that in 1994, roughly 50 percent of students correctly answered items involving number concepts compared with questions involving estimation, in which only 44 percent of students responded correctly. Notice that on an absolute scale, students appear to have made the most progress in items involving computation (6.8 percentage point increase) and number concepts (7.1 percentage points). Note that even relatively small differences are statistically significant because of the large samples. Within the category of number concepts, we see that students made the largest gain on items involving probability and statistics and the smallest relative gain on items involving measurement. Within computation, students made the most progress on questions involving fractions. Table 6 examines the item gains across student achievement level. Interestingly, we see the same pattern within each student ability grouping—the largest gains coming from computation and concepts and the smallest gains coming from problem-solving.

Table 7 shows similar item achievement trends for reading. The broadest categorization of reading items involves three question categories: those that require students to construct factual meaning, evaluative meaning and inferential meaning. We see that the gains were nearly identical across these categories. A more detailed categorization breaks down each of these broad categories into specific skills. Within evaluative meaning, questions asking students to determine the main idea or assess the author's viewpoint showed the largest gains while questions asking students about the style of a passage or to interpret non-literal language gained the least. Among inference questions, those asking students to infer the traits of characters or apply information showed the largest gains. Predict likely outcomes and infer feelings of characters showed the least improvement. Table 8 shows that all student ability groups showed similar patterns of responses – generally equal gains across the three broad categories, with somewhat smaller gains in determine factual information.

Table 9 shows ITBS achievement gains by item difficulty. Here item difficulty is defined in terms of the percentage of children correctly answering an item during the Spring norming sample of the exam conducted by the test publisher, Riverside, prior to the release of the exam. Therefore, these categories represent how difficult an average cross section of students in the U.S. found these questions. In math, we see the largest absolute gains among items of moderate difficulty and the largest relative gains among the most difficult items. In reading, the largest absolute as well as relative gains were for moderate difficulty items—i.e., those items where 41 to 70 percent of students in the nation answered correctly. This is contrary to the view that test-based accountability leads students and teachers to focus on the "low hanging fruit" – that is, questions on which it is easiest to make quick improvement.

Conclusions

As states seek to implement the mandates of NCLB, educators will increasingly face the task of interpreting test score gains in the context of test-based accountability. Perhaps the most important question is whether the test score gains are meaningful – that is, reflect an increase in some set of knowledge or skills. Because of differences in content and emphasis across exams, we would not expect gains on one test to be completely generalized to another test. By carefully examining the nature of the gains on a high-stakes exam—e.g., how achievement varied across item type and difficulty and to what extent changes in test completion and/or guessing explain the overall improvement—it is possible to obtain a better understanding of the performance changes under an accountability program. Item level analysis of test score gains in Chicago during the 1990s reveals several findings. First, while guessing increased following the introduction of accountability, it alone can only explain a small fraction of the observed test score changes. Second, the large observed math test score gains came disproportionately in the areas of computation and number concepts, areas that measure knowledge of basic skills more than complex thinking. In contrast, the improvements in reading appear to be spread relatively evenly across item types. Finally, the test score gains in Chicago were not disproportionately from easy questions, suggesting that teachers were not simply teaching skills that are most quickly mastered.

References

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.

Goodnough, A. 1999. "Answers Allegedly Supplied In Effort to Raise Test Scores. The New York Times, December 8.

Haney, W. 2000. "The Myth of the Texas Miracle in Education Reform." In L. McNeil (Chair), The New Discrimination: Creating and Recreating Discrimination in Public Schools, Symposium presented at the annual meeting of the American Educational Research Association, New Orleans, April 27.

Klein, S., L. Hamilton, D. McCaffrey, and B. Stecher. 2000. "What Do Test Scores in Texas Tell Us?" (IP-202) Santa Monica, CA: RAND.

Koretz, D.M. Forthcoming. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity." The Journal of Human Resources.

Koretz, D.M. 1988. "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction? American Educator, Summer, 12(2): 8-15, 46-52.

Koretz, D.M. 1992. "State and National Assessment." In M.C. Alkin, ed. Encyclopedia of Educational Research, Sixth Edition, Washington, D.C.: American Educational Research Association, 1262-1267.

Koretz, D.M. 1998. "Large-scale Portfolio Assessments in the US: Evidence Pertaining to the Quality of Measurement." In D. Koretz, A. Wolf, and P. Broadfoot ,eds., Records of Achievement. Special issue of Assessment In Education, 5(3): 309-334.

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.

Shavelson, R.J., and N.M. Webb. 1991. Generalizability Theory: A Primer. Newbury Park: Sage.

Shepard, L.A. 1988a. "Should Instruction Be Measurement-driven?: A Debate." Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Shepard, L.A. 1988b. "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. 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., and K.D. Dougherty. 1991. "Effects of High-stakes Testing on Instruction." 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, IL, April.

Stecher, B.M, and S.I. Barron. 1999. "Quadrennial Milepost Accountability Testing in Kentucky." CSE Technical Report No. 505. Los Angeles: Center for the Study of Evaluation, University of California.

Wilgoren, J. 2001. "Possible Cheating Scandal is Investigated in Michigan." The New York Times, June 9.

"Quality Counts 2002." *Education Week* 21(17): 74-77.

Tysome, T. (1994, August 19). Cheating purge: Inspectors out. Times Higher Education Supplement, p. 1.

Lindsay, D. (1996, October 2). Whodunit? Officials find thousands of erasures on standardized tests and suspect tampering. *Education Week*, 25-29.

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

Kolker, Claudia. (1999). "Texas Offers Hard Lessons on School Accountability." *Los Angeles Times*, April 14, 1999.

Hofkins, D. (1995, June 16). Cheating "rife" in national tests. Times Educational Supplement, p. 1.

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

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

Table 1: Descriptive Statistics on ITBS Completion Rates Before and After High-Stakes Testing

	Math			Reading		
	1994	1998	% Point Change	1994	1998	% Point Change
All Students						
Fraction of items completed	0.968	0.979	0.011	0.966	0.987	0.021
Fraction correct/fraction completed	0.466	0.520	0.054	0.508	0.558	0.050
No Blanks	0.577	0.632	0.055	0.773	0.834	0.061
No Final Blanks	0.722	0.817	0.095	0.857	0.917	0.060
Guessing	0.171	0.234	0.063	0.043	0.083	0.040
Low Ability Students (0-25 percentile)						
Fraction of items completed	0.960	0.973	0.013	0.956	0.973	0.017
Fraction correct/fraction completed	0.371	0.402	0.031	0.397	0.425	0.028
Guessing	0.148	0.227	0.079	0.034	0.086	0.052
Moderate Ability Students (26-50 percentile)						
Fraction of items completed	0.970	0.980	0.010	0.969	0.982	0.013
Fraction correct/fraction completed	0.483	0.516	0.033	0.532	0.556	0.024
Guessing	0.203	0.259	0.056	0.051	0.090	0.039
High Ability Students (51-99 percentile)						
Fraction items completed	0.981	0.986	0.005	0.984	0.991	0.007
Fraction correct/fraction completed	0.650	0.680	0.030	0.717	0.733	0.016
Guessing	0.173	0.212	0.039	0.049	0.069	0.020

Notes: Sample consists of all tested and included first-time eighth grade students in 1994 and 1998. The prior achievement categories are based on the average of math and reading score in 5th grade, with the percentiles referring to percentiles on a national distribution. Guessing is measured by a series of identical, incorrect items in the last 3 questions on the exam (e.g., AAA, BBB, CCC or DDD).

Table 2: Estimates of the achievement gain associated with greater student guessing

	OLS estimates of the number of questions answered on the ITBS	Estimated additional correct responses with random guessing	Achievement gain associated with additional correct response	Estimated gain associated with random guessing	Observed gain	Maximum fraction explained by guessing
	(1)	(2) = (1) *0.25	(3)	(4) = (2)*(3)	(5)	(6)=(4)/(5)
3rd Grade						
Reading	0.37	0.0925	0.154	0.0142	0.115	0.123
Math	1.03	0.2575	0.050	0.0129	0.186	0.069
6th Grade						
Reading	0.62	0.155	0.232	0.0360	0.359	0.100
Math	1.03	0.2575	0.073	0.0188	0.360	0.052
8th Grade						
Reading	0.80	0.20	0.252	0.0504	0.381	0.134
Math	1.84	0.46	0.072	0.0331	0.284	0.116

Notes: The sample includes all first-time students in these grades in 1994 and 1998. Control variables include race, gender, race*gender interactions, age, household composition, and an indicator of previous special education placement along with up to three years of prior reading and math achievement (linear, square and cubic terms). Missing test scores are set to zero and a variable is included indicating the score is missing. Robust standard errors that account for the correlation of errors within school are presented in parentheses.

Table 3: The Relationship between High-Stakes Testing and Exam Completion

	Dependent Variables			
	# blank items on the exam	**No Blanks**	**No Final Blanks**	**Any Guessing**
Model	Negative Binomial	Probit	Probit	Probit
Estimate shown	IRR	dF/dx	dF/dx	dF/dx
Reading (49 items)				
High-Stakes Effect	0.44	0.079	0.064	0.046
	(9.5)	(8.8)	(9.6)	(6.0)
Baseline Level	1.65	0.77	0.86	0.04
Math 1 – Concepts (56 items)				
High-Stakes Effect	0.47	0.065	0.046	0.026
	(9.2)	(8.0)	(9.5)	(4.5)
Baseline Level	1.26	0.79	0.88	0.06
Math 2 – Problem-Solving (36 items)				
High-Stakes Effect	0.60	0.021	0.015	0.004
	(5.1)	(4.0)	(5.4)	(1.2)
Baseline Level	0.26	0.91	0.96	0.08
Math 3 – Computation (43 items)				
High-Stakes Effect	0.48	0.113	0.099	0.041
	(11.0)	(9.9)	(10.4)	(5.4)
Baseline Level	2.86	0.67	0.76	0.06

Notes: The sample includes all tested and included first-time 8th graders. The coefficient estimate for high-stakes testing is for the 1998 cohort, and is comparable for the other cohorts. Guessing is measured by a series of identical, incorrect items in the last 3 questions on the exam (e.g., AAA, BBB, CCC or DDD). Coding errors include marking multiple answers for one questions, shading a response too lightly, or leaving any stray marks on that item. Coefficient estimates for negative binomial regressions are shown as incident rate ratios. Coefficient estimates for Probit models are shown as marginal effects evaluated at the mean. Robust t-statistics that account for the clustering of students within schools are shown in parenthesis (note: for the negative binomial regressions, these are actually z-statistics).

Table 4: The Relationship between High-Stakes Testing and ITBS Math Achievement

	Sample		
Dependent Variable	**3rd Grade**	**6th Grade**	**8th Grade**
Reading Total	0.173 (0.019)	0.212 (0.014)	0.184 (0.019)
Math Total	0.213 (0.021)	0.242 (0.017)	0.288 (0.024)
Math Section 1 (Math Concepts and Estimation)	0.186 (0.021)	0.188 (0.016)	0.257 (0.016)
Math Section 2 (Math Problems and Data Interpretation)	0.158 (0.019)	0.145 (0.013)	0.165 (0.013)
Math Section 3 (Math Computation)	0.252 (0.024)	0.342 *(0.024)*	0.409 (0.024)
Average 1994 Scores (in grade equivalents)			
Reading Total	3.04	5.88	7.70
Math Total	3.39	6.14	7.80
Math Section 1	3.41	6.12	7.69
Math Section 2	3.25	5.94	7.77
Math Section 3	3.51	6.37	7.93

Notes: Cells contain OLS estimates of the impact of high-stakes testing for the 1998 cohort. The outcome measures are standardized using the 1993 student-level mean and standard deviation. Robust standard errors that account for within school correlation of errors are shown in parentheses. Other variables included in the regressions but not shown here include the following: include race, gender, race*gender interactions, guardian, bilingual status, special education placement, prior math and reading achievement, school demographics (including enrollment, racial composition, percent free lunch, percent with limited English proficiency and mobility rate) and demographic characteristics of the student's home census tract (including median household income, crime rate, percent of residents who own their own homes, percent of female-headed household, mean education level, unemployment rate, percent below poverty, percent managers or professionals and percent who are living in the same house for five years). Prior achievement is measured by math and reading scores three years prior to the base year (i.e., at *t-3*). Missing test scores are imputed using other observable characteristics of the student and a variable is included indicating the score was missing. Second and third-order polynomials in prior achievement are included to account for any non-linear relationship between past and current test scores.

Table 5: The Relationship between HST and ITBS Math Achievement

	The proportion of students answering the type of item correctly in				
Item Type	**1994**	**1996**	**1998**	**% Point Gain (1994-1998)**	**% Gain (1994-1998)**
Number Concepts	0.497	0.521	0.568	0.071	0.142
CONC EQUATIONS & INEQUALITIES	0.527	0.553	0.600	0.073	0.138
CONC FDP	0.489	0.514	0.565	0.076	0.155
CONC GEOMETRY	0.541	0.569	0.614	0.073	0.136
CONC MEASUREMENT	0.388	0.399	0.424	0.035	0.091
CONC NUMERATION & OPERATIONS	0.535	0.558	0.607	0.073	0.136
CONC PROBABILITY & STATISTICS	0.387	0.414	0.473	0.086	0.222
Estimation	0.440	0.459	0.494	0.053	0.121
ESTI COMPENSATION	0.341	0.349	0.366	0.025	0.075
ESTI ORDER OF MAGNITUDE	0.531	0.553	0.590	0.059	0.112
ESTI STANDARD ROUNDING	0.496	0.521	0.569	0.074	0.149
Problem-Solving	0.465	0.479	0.508	0.043	0.093
PROB MULTIPLE STEP	0.421	0.435	0.465	0.044	0.104
PROB PROBLEM SOLVING STRATEGIES	0.383	0.399	0.421	0.038	0.098
PROB SINGLE STEP	0.599	0.612	0.645	0.046	0.076
Data Interpretation	0.478	0.500	0.534	0.055	0.115
DATA COMPARE QUANTILES	0.434	0.454	0.488	0.053	0.123
DATA INTERPRET RELATIONSHIPS & TRENDS	0.464	0.485	0.514	0.050	0.107
DATA READ AMOUNTS	0.554	0.578	0.622	0.068	0.124
Computation	0.516	0.529	0.584	0.068	0.132
COMP DECIMALS_ADD	0.506	0.521	0.587	0.082	0.162
COMP DECIMALS_DIVIDE	0.293	0.301	0.351	0.058	0.198
COMP DECIMALS_MULTIPLY	0.414	0.415	0.459	0.045	0.110
COMP DECIMALS_SUBTRACT	0.539	0.563	0.612	0.073	0.135
COMP FRACTIONS_ADD	0.447	0.460	0.537	0.090	0.202
COMP FRACTIONS_DIVIDE	0.364	0.374	0.471	0.107	0.294
COMP FRACTIONS_MULTIPLY	0.346	0.353	0.420	0.073	0.212
COMP FRACTIONS_SUBTRACT	0.330	0.341	0.421	0.091	0.277
COMP WHOLE NUMBERS_ADD	0.707	0.717	0.756	0.049	0.070
COMP WHOLE NUMBERS_DIVIDE	0.546	0.554	0.599	0.054	0.098
COMP WHOLE NUMBERS_MULTIPLY	0.528	0.541	0.594	0.067	0.126
COMP WHOLE NUMBERS_SUBTRACT	0.628	0.642	0.685	0.057	0.091

Notes: Sample includes students in grades three, six and eight for the first time who were tested and included for reporting purposes.

Table 6: The Relationship between HST and ITBS Math Achievement by Student Prior Achievement

	The proportion of students answering the type of item correctly in			**% Point Gain (1994-1998)**	**% Gain (1994-1998)**
All students	**1994**	**1996**	**1998**		
COMPUTATION	0.516	0.529	0.584	0.068	0.132
CONCEPTS	0.497	0.521	0.568	0.071	0.142
DATA INTERPRETATION	0.478	0.500	0.534	0.055	0.115
ESTIMATION	0.440	0.459	0.494	0.053	0.121
PROBLEM SOLVING	0.465	0.479	0.508	0.043	0.093
Low Ability Students (0-25 percentile)					
COMPUTATION	0.396	0.405	0.452	0.056	0.142
CONCEPTS	0.371	0.389	0.423	0.052	0.139
DATA INTERPRETATION	0.359	0.369	0.397	0.038	0.107
ESTIMATION	0.344	0.352	0.383	0.039	0.113
PROBLEM SOLVING	0.340	0.347	0.368	0.028	0.083
Moderate Ability Students (26-50 percentile)					
COMPUTATION	0.531	0.538	0.588	0.057	0.107
CONCEPTS	0.509	0.526	0.568	0.059	0.116
DATA INTERPRETATION	0.495	0.511	0.538	0.043	0.087
ESTIMATION	0.449	0.463	0.494	0.044	0.098
PROBLEM SOLVING	0.469	0.475	0.496	0.027	0.057
High Ability Students (51-99 percentile)					
COMPUTATION	0.698	0.701	0.741	0.043	0.062
CONCEPTS	0.702	0.712	0.750	0.048	0.068
DATA INTERPRETATION	0.666	0.679	0.701	0.035	0.053
ESTIMATION	0.601	0.617	0.636	0.035	0.058
PROBLEM SOLVING	0.673	0.681	0.696	0.023	0.034

Notes: Sample includes students in grades three, six and eight for the first time who were tested and included for reporting purposes. Compositional changes (i.e., the increase in prior achievement levels from 1994 to 1998) is the reason that the trends for all students are not simply averages of those for all the three prior achievement groups.

Table 7: The Relationship between HST and ITBS Reading Achievement

	The proportion of students answering the type of item correctly in			% Point Gain (1994-1998)	% Gain (1994-1998)
Item Type	1994	1996	1998		
CONSTRUCT EVALUATIVE MEANING	0.491	0.510	0.545	0.054	0.110
EVA AUTHOR'S PURPOSE	0.418	0.435	0.473	0.055	0.132
EVA AUTHOR'S VIEWPOINT	0.617	0.643	0.681	0.064	0.104
EVA DETERMINE MAIN IDEA	0.510	0.527	0.572	0.063	0.123
EVA INTERPRET NONLITERAL LANGUAGE	0.477	0.492	0.520	0.043	0.090
EVA STRUCTURE	0.357	0.378	0.409	0.053	0.148
EVA STYLE	0.566	0.582	0.600	0.033	0.059
CONSTRUCT FACTUAL MEANING	0.456	0.471	0.507	0.051	0.112
FAC LITERAL MEANING OF WORDS	0.385	0.400	0.430	0.045	0.116
FAC UNDERSTAND FACTUAL INFORMATION	0.469	0.484	0.521	0.053	0.112
CONSTRUCT INFERENTIAL MEANING	0.490	0.507	0.544	0.054	0.110
INF APPLY INFORMATION	0.488	0.517	0.555	0.067	0.136
INF DRAW CONCLUSIONS	0.475	0.490	0.524	0.049	0.103
INF INFER FEELINGS OF CHARACTERS	0.601	0.629	0.648	0.047	0.079
INF INFER MOTIVES OF CHARACTERS	0.495	0.505	0.552	0.058	0.116
INF INFER TRAITS OF CHARACTERS	0.584	0.618	0.674	0.091	0.155
INF PREDICT LIKELY OUTCOMES	0.440	0.442	0.470	0.030	0.068
INF REPRESENT INFORMATION IN NEW FORM	0.418	0.445	0.486	0.068	0.163
INF APPLY INFORMATION	0.488	0.517	0.555	0.067	0.136

Notes: Sample includes students in grades three, six and eight for the first time who were tested and included for reporting purposes.

Table 8: The Relationship between HST and ITBS Reading Achievement by Student Prior Achievement

	The proportion of students answering the type of item correctly in			% Point Gain (1994-1998)	% Gain (1994-1998)
	1994	1996	1998		
Al Students					
CONSTRUCT EVALUATIVE MEANING	0.491	0.510	0.545	0.054	0.110
CONSTRUCT FACTUAL MEANING	0.456	0.471	0.507	0.051	0.112
CONSTRUCT INFERENTIAL MEANING	0.490	0.507	0.544	0.054	0.110
Low Ability Students (0-25 percentile)					
CONSTRUCT EVALUATIVE MEANING	0.368	0.375	0.408	0.040	0.108
CONSTRUCT FACTUAL MEANING	0.330	0.333	0.360	0.030	0.089
CONSTRUCT INFERENTIAL MEANING	0.370	0.375	0.409	0.039	0.104
Moderate Ability Students (26-50 percentile)					
CONSTRUCT EVALUATIVE MEANING	0.510	0.520	0.551	0.041	0.080
CONSTRUCT FACTUAL MEANING	0.459	0.468	0.497	0.038	0.084
CONSTRUCT INFERENTIAL MEANING	0.494	0.509	0.540	0.046	0.093
High Ability Students (51-99 percentile)					
CONSTRUCT EVALUATIVE MEANING	0.696	0.704	0.721	0.025	0.036
CONSTRUCT FACTUAL MEANING	0.662	0.674	0.692	0.030	0.046
CONSTRUCT INFERENTIAL MEANING	0.683	0.699	0.716	0.033	0.048

Notes: Sample includes students in grades three, six and eight for the first time who were tested and included for reporting purposes. Compositional changes (i.e., the increase in prior achievement levels from 1994 to 1998) is the reason that the trends for all students are not simply averages of those for all the three prior achievement groups.

Table 9: The Relationship between HST and ITBS Reading Achievement

	The proportion of students answering the type of item correctly in			% Point Gain (1994-1998)	% Gain (1994-1998)
Item Type	1994	1996	1998		
Math					
21-30 percent correct	0.245	0.259	0.299	0.054	0.221
31-40 percent correct	0.313	0.326	0.360	0.048	0.153
41-50 percent correct	0.363	0.380	0.428	0.065	0.179
51-60 percent correct	0.448	0.467	0.511	0.063	0.141
61-70 percent correct	0.528	0.550	0.600	0.071	0.135
71-80 percent correct	0.648	0.666	0.708	0.060	0.093
81-90 percent correct	0.748	0.762	0.800	0.052	0.070
91-100 percent correct	0.803	0.815	0.849	0.046	0.057
Reading					
21-30 percent correct	0.213	0.210	0.219	0.006	0.030
31-40 percent correct	0.266	0.275	0.303	0.037	0.140
41-50 percent correct	0.357	0.373	0.413	0.056	0.156
51-60 percent correct	0.425	0.446	0.484	0.059	0.140
61-70 percent correct	0.525	0.543	0.584	0.059	0.112
71-80 percent correct	0.617	0.637	0.670	0.054	0.087
81-90 percent correct	0.753	0.770	0.801	0.048	0.064
91-100 percent correct	0.828	0.826	0.859	0.032	0.039

Notes: Sample includes students in grades three, six and eight for the first time who were tested and included for reporting purposes.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034922

I. DOCUMENT IDENTIFICATION:

Title: Test-Based Accountability and Student Achievement Gains: Theory and Evidence

Author(s): Brian A. Jacob

Corporate Source:

Publication Date: June 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature:

Printed Name/Position/Title: Brian Jacob

Organization/Address: Kennedy School of Govt.
79 JFK St. Cambridge, MA 02138

Telephone: 617 384-7966

FAX:

E-Mail Address:

Date: 5/10/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse: **University of Maryland**
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions