"INFLATED TEST SCORE GAINS": IS IT OLD NORMS OR TEACHING THE TEST?

CSE Technical Report 307

Lorrie A. Shepard

University of Colorado

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

January, 1990

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This paper was written as part of a research project sponsored by the UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST). It was prepared for presentation at the Annual Meeting of the American Educational Research Association, San Francisco, March, 1989.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Introduction

Cannell's 1987 report attacked the credibility and integrity of nationally normed standardized achievement test results. According to his survey, all 50 states claim to be above the national average and an estimated 70% of students nationally are told they are performing above average. Cannell found these results illogical and inconsistent with other indicators of educational quality. Although he had heard the counterexplanation that "high scores reflect improved achievement levels," he argued that inaccurate initial norms and "teaching the test" were more likely causes of high scores.

Responses to Cannell from educational policymakers and from test publishers were of three types: (a) his data are wrong (or he doesn't understand statistics); (b) his negative inferences are wrong, high achievement scores are real; or (c) he's right, test scores are very likely inflated by factors such as outdated norms and too much familiarity with the tests. Linn et al. (1989) addressed the validity of the first two rebuttals. His analysis provided more exhaustive consideration of subject areas and grade levels, more statistically defensible treatment of reported test score distributions, and a more representative sample of school district data. Nonetheless, he confirmed Cannell's basic conclusion. After considering reported results from the 35 states with nationally normative comparisons, he stated that "the overall percent of students above the national median is greater than 50 in all of the elementary grades in both reading and mathematics for each of the three years studied" (p. 8). Use of the median rather than mean precludes esoteric discussion about skewed distributions. By definition, 50% of students should be on each side of the median. Thus, only two assertions are possible: Either achievement has gone up since the base year, or something is amiss.

Linn also presents data that contradict the claim that all of the apparent gains are real. National trend data provided by the National Assessment of Educational Progress (NAEP) document gains in achievement that are much more modest than the dramatic gains reported by many state assessments and by publishers for their normative samples. Koretz's (1986, 1987) studies for the Congressional Budget Office of several large-scale databases confirm that achievement is improving nationally, but Koretz (1988) concluded that the gains reported by standardized tests are exaggerated.

These comparisons to more credible national data both support and contradict Cannell's claims. Yes, achievement gains reported to the public based on standardized achievement tests appear to be exaggerated. But it is also apparent that the norms themselves may be inflated; the steep gains from the 1970s to the 1980s norm groups could be caused in part by oversubscription of prior users in the recent norm groups (see Table 1, Phillips & Finn, 1988) and similarities between old and new forms of the tests. This would mean, contrary to Cannell's accusation of collusion and misrepresentation by publishers to make schools look good, that the revised norms actually could have set too high a standard of comparison in the base year. Furthermore, if up-to-date norms carried forward this intrinsic response bias, they would continue to be too high, not too low. This is an extremely important point because it bears on the validity of alternative solutions to the problems raised by Cannell.

If outdated norms are seen as the central problem, then annual norms are the answer. Indeed, annual norms and educating the public about the "time-bound nature of norms" (Williams, 1988) have been the primary responses by test publishers and state testing directors in our survey. If, however, the problem of spuriously high test scores is a result of too much teaching the test in the face of too much accountability pressure, then annual norms will contribute to the problem by

creating a standard that is more and more unattainable by legitimate teaching methods. This tension or dilemma is the focus of the paper.

This report presents (a) an overview of the explanations given for spurious test score gains and (b) an encapsulated summary of findings from our survey of state testing directors regarding the narrowing of curriculum and teaching the test. Then I return to the dilemma posed by the effect of teaching the test on the norms themselves and consider what solutions should be pursued if test familiarity is seen as the primary problem, rather than outdated norms.

Explanations for Spurious Test Score Gains

Cannell first released his report in November of 1987; the Summer 1988 issue of *Educational Measurement: Issues and Practice* was a special issue devoted to Cannell's findings with commentaries by researchers representing the U.S. Department of Education and each of the major test publishing firms. Table 1 provides a summary of the elements in those responses that specifically address the possible explanations for inflated scores. Explanations 1 and 4 pertain to norms. Explanations 2, 3, and 5 refer to aspects of teaching the test. Note that Drahozal and Frisbie (1988) and Stonehill (1988) speculated about the type of bias that would have to occur for non-representative norms to lead to an overestimate of student achievement. In contrast, Phillips and Finn (1988) and Lenke and Keene (1988) considered the more realistic possibility that normative samples become biased by the greater participation rates of user districts, thus leading to spuriously high norms and an underestimate of achievement for naive test takers. Lenke and Keene provided direct evidence that user norms are inflated but did not apply these findings to argue against the validity of annual norms (which publishers would most likely construct from user data). Instead they argued against annual norms as a "moving target" that would preclude evaluation of change over time. Three respondents suggested annual user norms as a corrective to outdated norms. Three respondents suggested fresh tests or test security as the remedy for teaching the test. Only one author compared the two problems and their respective solutions: Anticipating the theme of this paper, Qualls-Payne (1988) commented, "If new forms of achievement tests are developed each year, thereby increasing test security, the need for annual norms diminishes significantly" (p. 22).

Various Meanings of Teaching the Test

The phrase "teaching the test" is evocative, but in fact it has too many meanings to be directly useful. Although it has a negative connotation for most members of the public, many educators take it to mean teaching to the domain of knowledge represented by the test. In framing our interview questions with state testing directors or their representatives, we avoided the pejorative phrase with multiple interpretations. Instead, we asked about a wide range of policies and practices, beginning with the uses for the test data, the process of test selection, time spent on teaching the test objectives, and test preparation efforts.

It is commonly understood that one of the salient characteristics of the educational reform movement of the 1980s has been high-stakes testing. Popham (1987) used the term "high stakes" to refer to tests with severe consequences for individual pupils, such as non-promotion, and those used to rank schools and districts in the media. The latter characterization clearly applies to 40 of the 50 states. Only 4 states conduct no state testing, nor do they aggregate local district results (Montana, Nebraska, Ohio, and Vermont); 2 states, Oregon and Wyoming, collect state data on a sampling basis in a way that does not put the spotlight on local districts. Wisconsin and North Dakota report state results collected from districts on

Table 1. Explanations for Spuriously High Achievement Scores From Responses to the Cannell Report

	Phillips & Finn U.S. Dept. of Educ.	Drahozal & Frisbie Riverside Publishing	Lenke & Keene Psychological Corp.	Williams CTB/McGraw-Hill	Qualls-Payne SRA	Stonehill U.S. Dept. of Educ.
1.	Non-representativeness of national norms; overrepresentation of test users leads to spuriously high norms.	Non-representativeness of national norms: underrepresentation of high scoring districts leads to spuriously low norms.	Users outperform non-users in sample.			Non-representative norms would be "too easy" if they rely too much on compensatory education students
2.	Curriculum alignment in test selection gives an advantage over norming condition.	Curriculum alignment will lead to overestimate of pupil standing (see 5).	Test users selecting test matched to curriculum have an advantage.			
3.	High stakes pressure creates more motivation than in norming condition.	Job retention and salary increases tied to scores (see 5).				
4.	Outdated norms. Solutions = annual user-norms if representative.	Recency of norms. 1970s norms are "softer" than 1980s norms. Planning for annual norms.	Achievement is going up. But changing norms too often would create a "moving target."	Norming cycles are well known; more above median scores are valid measures of rising achievement. Annual norms for users.	User-based norms to monitor achievement trends and signal need for renorming.	
5.	Teaching the test (rather than the test objectives). Solution = fresh tests more often.	Teaching the test. Practice and narrowing of the curriculum. Solution = test security.			New test each year would reduce need for annual norms.	
6.	Non-comparability of samples; more holding out of low scoring students in user sample than in norming sample.	Difference between tested and total enrolled population.	Handicapped and limited English speakers may not be excluded by same guidelines.			
7.		Adequacy of expectations based on socio-economic factors and expenditures.				
8.		NRTs originally intended to evaluate pupil scores.				
		Accuracy of comparisons pupil vs. district averages.	Interpreting group performance relative to national pupil norms.			

Note. Some authors' responses to Cannell disputed his facts and statistics or argues that the test score gains were real rather than spurious. Such responses are not included in Table 1 which summarizes only the explanations given for spuriously high scores.

a voluntary basis. Two additional states were rated as relatively low-stakes by their test coordinators;[1] in these states, for example, test results are not typically page-one news nor are district rank-orderings published. The testing directors in the 10 states without high-stakes state tests were careful to point out that their comments did not necessarily apply to individual districts within their state where public attention to test scores might be extremely intense.

The most pervasive source of high-stakes pressure identified by respondents was media coverage. Presentation of test results to the state board is a media event. Each local paper then runs its own story on the health of public education and ranks the districts within its jurisdiction. Other uses of test results that would give them extremely high importance where they occur were reportedly infrequent. For example, many respondents had heard talk of superintendents or principals who were fired because they had been unable to raise test scores satisfactorily. Though the talk was widespread, contributing in many cases to reported principal anxiety, the instances were rare and unverified in all but a few cases. Only a few states have financial incentive programs, in which there is some financial reward to schools, districts, or teachers that derives from raised test scores. Another small number of states have placed districts in receivership, based on low test scores and other factors. Although none of the states reported the coincidence of all of these high-stakes pressures, intense media coverage and scrutiny from the legislature alone were sufficient for many to rate test score results as "very important" or "extremely important." The other factors appeared to contribute to the sense of urgency or pressure associated with test scores even if they directly affected only a small portion of educators in the state.

High stakes do not necessarily mean invalid test scores, but they clearly alter the context of testing, as suggested by Phillips and Finn (1988). Furthermore, intense pressure on educators to improve scores sets the stage and increases the incentives for the various types of teaching the test efforts discussed in the following sections.

Test-Curriculum Alignment

Without question, published norm-referenced tests are selected to achieve the best match possible between the test content and the state's curriculum. The following interview segment typifies the process described by state directors in response to the question, "Who selected the standardized test being used?"

> Committees of teachers are set up by grade level so that a group of third-grade teachers would be reviewing tests appropriate to the third grade, and then would begin making recommendations as to which test is better in content, format, technical characteristics, and so forth. We also add to that list of teachers groups of technical specialists who look at things like norms and so forth, adequacy of reporting and scaling and so forth. We also add another committee comprised mostly of persons that would be curriculum specialists at the central office level. And these three groups make independent recommendations. HAVE THERE BEEN EFFORTS TO ASSURE THAT THE CURRICULUM AND THE TEST ARE ALIGNED? Absolutely. That's what each of these three groups does. The teachers look at a lot of things like formatting and carefulness of construction and look for item bias, those kinds of things. But the key thing that they

[1] Our interviews were conducted under the agreement with respondents that states would not be anonymous when citing matters of fact regarding the testing program or policies that would be quoted directly from published materials. However, when respondents were asked to state an opinion or perception, they, and hence their states, would be anonymous.

look for is alignment with curriculum. If the test is not aligned with our curriculum, it just gets discarded immediately.

The few states with customized tests or home-made tests linked to national norms are able, of course, to achieve even closer alignment between the state's curriculum outline and test content because they are not constrained to select from existing tests.

It is also evident that test-curriculum alignment is a reciprocal process. That is, once the test that best fits the curriculum is chosen, the practiced curriculum is adjusted further in response to the test. Many directors emphasized that this was, in fact, the conscious purpose of the testing program, to ensure that essential skills are taught. Item analysis data are usually provided and districts are encouraged to look for areas of weakness that require greater instructional effort. Counterexamples were extremely rare; for example, we were told by one respondent that districts are told not to worry about subpart scores where they do poorly if that element is not emphasized in their local curriculum or is taught at a later time.

When asked, "Do you think that teachers spend more time teaching the specific objectives on the test(s) than they would if the tests were not required?," the answer from the 40 high-stakes states was, nearly unanimously, "Yes." The majority of respondents went on to describe the positive aspects of this more focused instruction: "Surely there is some influence of the content of the test on instruction. That's the intentional and good part of the testing, probably." And in another state, "I can only tell you that the people I've talked to, and it is certainly not a representative sample, have indicated that in fact the presence of the test is forcing attention to the essential skills that had been identified." Other respondents (representing about one-third of the high-stakes states) also said that teachers were spending more time teaching the specific objectives on the test, but they cast their answer in a negative way: "Yes. There is some definite evidence to that effect. I don't know that I should even say very much about that. There are some real potential problems there...Basically the tests do drive the curriculum."

The follow-up question, "To what extent do you think important objectives are given less time or emphasis because they are not included in the test?," elicited a less uniform response, but answers were consistent with the positive or negative valence to the preceding answer. For example, those who believed that focusing instruction was a positive effect of testing gave answers such as the following:

> "Yes, that happens, but in a minority of our schools."
> "They would teach the essential competencies even without the [norm referenced test]."
>
> "Until the students master the basic skills their experiences in other areas are limited or non-functioning anyway."
>
> "There's some tendency to narrow, but the community keeps the pressure on for gifted education."

Those who expressed more concern about the narrowing of instruction gave answers such as:

> The answer is yes, but I have no idea. I'm not close enough to any data that would give me a clue on percentage. I certainly feel comfortable saying yes, that I think there has been a decreased emphasis. WHAT KINDS OF THINGS ARE OMITTED OR DEEMPHASIZED? I think it occurs

two ways. One, within the subject, some of the higher level objectives suffer—that is, other than reading and math.

Test selection to match curriculum and subsequent shaping of curriculum to conform to the test are not regarded as illegitimate practices. For decades, it has been standard advice in measurement textbooks to select standardized tests on the basis of technical adequacy and congruence with local curricula. Aligning curriculum to follow the test can be defended in the general spirit of teaching to agreed upon goals; whether particular instances of this practice are defensible depends on the breadth of the test content and how extensively the tested objectives take over instruction. Although measurement-driven instruction may not be desirable if one rejects an assembly-line conception of learning (Bracey, 1987; Shepard, 1988), it is not patently unethical to teach to the test objectives.

In one sense it can be said that test-curriculum alignment does not lead to spurious test score gains. Students can be said to have learned more of the specified objectives. Narrowing of curriculum does, however, alter the meaning of normative comparisons. The original standardization sample did not have the benefit of such focused instruction. Students in the norming sample apparently were learning the tested content and other things as well when they took the unannounced test. One way of thinking about the change in the meaning of norms is to recall the old anchor study where national probability samples of students were used to equate all of the standardized tests to each other (Jaeger, 1973). When all tests are administered in naive conditions (i.e., where curricula have not been aligned), then the equating answers the question "How would students who performed at percentile X on test A, do on test B?" As soon as schools begin to tailor instruction to a particular test, these equivalences no longer hold. As far as the public meaning of test scores is concerned, however, there is an implicit assumption made that these equivalences hold true. For example, if the average student in your local district were scoring at the 60th percentile on the CAT, you would want to be able to assume that the district's performance would be roughly the same on the ITBS. More to the point, consider the political ethos associated with educational reform. When politicians learn that U.S. students do poorly on international achievement comparisons and install testing programs to increase achievement, they wish to assume that rising local scores are evidence that the achievement deficit has been remedied. But once curriculum has been aligned to the local test, there is no guarantee that apparent gains generalize to other tests.

Note that the provision of annual user norms moves in the opposite direction of the anchor study. The notion of naive test takers is abandoned and each test then develops its own population of users. A local district that uses a test by maintaining a broad curricular focus beyond the test domain would be a disadvantage in such comparisons.

Test Preparation

Our questions about test preparation were intended to encompass a range of activities including content review, advice about test-wiseness skills, and practice with unfamiliar formats, as well as the more questionable practices that Phillips and Finn (1988) had in mind when they referred to teaching the test as distinct from teaching test objectives. Respondents' descriptions of typical test preparation practices most often began with advice to students to "get a good night's sleep" the night before the test. Next most frequent was the response that districts use the standard materials provided by test publishers; especially, children in Grades 1, 2, and 3, and sometimes 4, are administered a formal practice test to acquaint them with test format demands. These materials are provided by the publishers, and unlike many practices that depart from the conditions of the standardization study, were a part of the normative test administration.

Most states do not provide materials for test preparation (beyond those available from the publisher), nor do they provide guidelines as to what constitutes appropriate test preparation. Several of the states with state-developed criterion-referenced tests distribute detailed item specifications to encourage teaching to the test objectives and distribute old forms of the state test to be used for student practice and remediation. Respondents at the state level were generally unaware of the extent to which local schools and districts engaged in content review or provided additional format practice for their students.

When asked what they had observed as extreme instances of test preparation, responses included:

> "Some districts have picked up on *Scoring High* [Cohen & Forman, 1987], which is not covered under our test security rules."
>
> "Once in a great while we find that people are using materials identical to our test."
>
> "Some districts have developed their own practice tests and have a timeline for covering each objective."
>
> "They have courses designed to prepare for the [high school] tests."
>
> "Pep rallies [are held] prior to test week to psych kids up to do well on the test."

One-time practice with test format, especially when such activities are consistent with standardization procedures, is not the cause of inflated test scores. However, repeated practice or instruction geared to the format of the test rather than the content domain can increase scores without increasing achievement. For example, Mehrens and Kaminski (1988) conducted a content analysis of the *Scoring High* test preparation materials published by Random House. They concluded that the materials were so similar to the test that practice with the *Scoring High* (CAT) is equivalent to giving the parallel form of the test as a practice test and explaining all the answer choices to the students. Although the latter would be clearly unethical, many educators purchase *Scoring High* without confronting any ethical issues because it is sold as instructional or review material.

Test Security and Test Familiarity

In two states security measures associated with the norm-referenced testing program were described by the state-level directors as lax; specifically, local schools were allowed to store testing materials in the school from one year to the next. These were the exceptions, however. The great majority of states described extensive security procedures intended to limit the exposure of test materials in the schools and to keep account of test booklets. The following excerpt from the Rhode Island *Testing Coordinator's Handbook* (1988) is an example of typical security precautions:

> 1. Store materials in rooms or cabinets that are locked, and that are not readily accessible to large numbers of other people.
>
> 2. Check all materials as you receive them to verify counts; have counts verified again when material are returned for storage.
>
> 3. Keep all extra test materials in the secure location when they are not in use. (p. 8)

We also asked state directors about their experiences with cheating and about procedures they used to detect anomalous results. Cheating had been exposed in several states but it was generally believed to occur in a very tiny percentage of schools (1% to 3%). Only California has in place a computer-scanning procedure to detect significant numbers of erasures signaling that teachers might have redone answer sheets after they were turned in by students. Using this procedure, the California Assessment Program announced last September the names of 40 elementary schools (among 5000) suspected of cheating on their 1985-86 tests (Woo, 1988). A number of states use computer-assisted or informal means to check for extraordinary gains from one year to the next and then inspect the materials to see if there is any evidence of tampering. The great majority of states do not have procedures to detect anomalous results. On rare occasions they receive phone calls from a parent or educator in a neighboring district who complains about practices such as distributing a dittoed version of the test the day before for practice or helping students during the test administration. Follow-up investigation may be handled by the state or the district; most often the test is readministered when an invalid administration occurs.

Although test materials are kept under lock and key and reported instances of cheating are rare, typical norm-referenced testing practices do not conform to the type of rigid security associated with programs such as college admissions testing. With some help from counselors, standardized tests are usually administered by classroom teachers. The same form of the test is administered year after year. Table 2 provides a summary of both norm-referenced and criterion-referenced testing programs with an indication as to how long the identical test has been used. Given that publishers follow a cycle of test revision every seven or eight years, it is not surprising that a few of the testing programs have had their tests in place for six or seven years.

We speculated that test familiarity might allow teachers to improve the performance of their students innocently without consciously deciding to cheat by photocopying the test. For example, suppose a teacher couldn't help but remember several of the vocabulary items on the test or the teacher chose to do a science unit on one of the animals discussed in a text reading passage. Perhaps the teacher was distressed during the test last year when his third graders were asked to do money problems in a format they had never seen before, so the teacher decided to use examples of that format from that time on. To assess how much impact test familiarity could have on scores, we used published norm tables for Grades 3 and 6 for two of the most prevalently used tests, the Stanford Achievement Test (SAT) and the California Achievement Tests (CAT), and looked up the conversions of number-right scores to percentile ranks. At the median in reading, language, and mathematics, one additional item correct translated into a percentile gain from 2 to 7 points. This means that teachers could teach relatively innocently to just a few items and raise achievement points by several percentile points. For example, on the CAT, Form E, the vocabulary subtest constitutes half of the total reading score. At third grade, a student at the 49th percentile would, with one more item correct on the vocabulary subtest, increase to the 54th percentile. Suppose that half of the class already knows the vocabulary words the teacher has remembered (or would know them in the ordinary course of instruction), then the teacher only has to be sure that the rest of the class learns two vocabulary items to increase the class standing on the vocabulary subtest by five percentile points.

The old complaint against norm-referenced tests was that they were insensitive to instruction. They were constructed to represent relatively broad content domains; items were thought of as samples from this broad domain. It would take an enormous amount of instruction aimed at the full domain to move the class average by a single item. Our examples from the norms tables illustrate, however,

Table 2 (DRAFT). Teacher Familiarity With Specific Test Items in State-Testing Programs

State	Same Test	Since	Given by Teacher	Notes
Alabama				
SAT/OLSAT	Yes	1984	Yes	Use alternate form for teacher review.
Alaska				
Various	Yes		Yes	Local district choice of tests; variable number of years in use.
Arizona				
ITBS/TASK	Yes	1986	Yes	
Arkansas				
State test	Yes	198___	Yes	
MAT-6	Yes	1985-86	Yes	Becoming familiar, we just changed from SRA.
California				*30-40 forms reused; †grade 3, 1980; grade 6, 1982; grade 8,
State test	Yes*	Variable†	Yes	1984; grade 12, 1987 (1976).
Colorado				
ITBS/TAP	Yes	1986	Yes	
Connecticut	Once or			
State test	twice only		Yes	Form A in schools for make-up.
Delaware				
CTBS	Yes	1986	Yes	
Florida				
State test				Multiple forms; different grades, different subjects each year
Georgia				
ITBS/TAP	Yes	1986	Yes	
Hawaii				
SAT	Yes	1986	Yes	
Idaho				
ITBS/TAP	Yes	1986-87	Yes	Grades 6 and 8, 1988; grade 11, 1987.
Illinois				
State test		1987-88	Yes	Rotating items.
Indiana				
Customized		1987-88	Yes	First year of program.
CAT+				Test of cognitive skills will be the same each year.
Iowa				
ITBS	Yes	1985-86	Yes	
Kansas				
State test				
Kentucky				
Customized	Yes	1981	Yes	
Louisiana				
CAT	Yes	1987-88	Yes	Switched tests this year.
Maine				
State test			Yes	Matrix sampling; 33% turnover items/year.
Maryland				
CAT	Yes	1981-82	Yes	
Massachusetts				
State test			Yes	3,600 items; replace 20-30%.
Michigan				
State test	Yes	1980	Yes	

Minnesota State test				Planning to change test items.
Mississippi SAT	Yes	1981	Yes	Not administered every year (?).
Missouri State test		1987-88	Yes	New forms with rotating items.
Nevada SAT	Yes	1984-85	Yes	Grade 9, 1987.
New Hampshire CAT	Yes	1985	Yes	Early fall testing means teachers are not identifying with results.
New Jersey State MC test			Yes	Old versions are used for remediation.
New Mexico CTBS	Yes	1981	Yes	
New York State test			Yes	Elementary test new every 3 years. High school new every year.
North Carolina CAT	Yes	1986	Yes	
North Dakota SRA/ITBS	Yes		Yes	Compilation of local norm-referenced tests.
Oklahoma MAT-6	Yes	1986	Yes	
Oregon State test			Yes	New each year; sample of schools.
Pennsylvania State test			Yes	Some old and new items each year.
Rhode Island MAT-6	Yes	1986	Yes	
South Carolina CTBS	Yes	1983	Yes	
South Dakota SAT/TASK	Yes	1985	Yes	
Tennessee SAT/TASK	Yes	1985	Yes	
Texas State test			Yes	As much as 50% same items.
Utah CTBS	Yes	1984		
Virginia STA/ITBS	Yes	1988	Yes	
Washington MAT	Yes	1985	Yes	
West Virginia CTBS	Yes	1984-85	Yes	
Wisconsin CTBS	Yes	1982	Yes	Phasing out after 1988.
Wyoming Concurrent NAEP				

Note. Follow-up phone calls are scheduled to confirm some of the entries in Table 2. Please report errors to the author.

that teaching to specific items is enormously more efficient. In this sense, norm-referenced tests are quite sensitive or vulnerable to teaching to specific items.

The Solution Should Be Fresh Tests, Not Annual Norms

Interview data cannot support a calculation that would sort out precisely how much of apparent test score gains are real and how much are spurious. Our data do suggest that the conditions for inflated results exist—to a marked degree in some cases. Forty of the 50 states administer high-stakes state testing programs that place some amount of pressure on teachers, principals, and superintendents to raise scores. There is substantial documentation of test-curriculum alignment. Practices that can be described as teaching the test rather than the test objectives exist to an unknown degree in every state. Each of these factors will affect the validity of local scores and also will distort the meaning of annual user norms.

When Phillips and Finn (1988) discussed annual user norms as a solution to outdated norms, they were very clear that these norms should be representative of the national population; presumably they were concerned that the sample not be biased with respect to demographic characteristics. Thus far, however, the discussion of annual norms has not confronted the issue of what it would mean to adopt a conception of a norming population where everyone is teaching the same test. Consider what it would mean to try to interpret relative standing in a population of users. In any normative comparison, a district is at a disadvantage if it plays fair by teaching to a broad curricular domain and avoiding more than one-time practice on test format. This disadvantage would be exaggerated, however, in a comparison among users rather than in a comparison in which each is compared to the naive norming sample. There is no way to assure that users in the norm group will teach to the test to the same degree, not that all will avoid unethical practices. Sources of invalid gains would then be built into the norm. The standard of comparison based on user norms would be spurious and inflationary.

Conclusion

In this study we have been concerned primarily with what test-curriculum alignment and teaching the test might do to the meaning of scores. There is ample evidence here and elsewhere, however, that these practices harm instruction and learning as well. For example, Darling-Hammond and Wise (1985) found that teachers abandon the use of essay tests because they are inefficient in preparing students for multiple-choice tests. In the early childhood field the rising number of kindergarten retentions is associated not just with direct kindergarten promotion tests, as in the Georgia example, but with concerns about protecting the school's performance on standardized tests as remote as third grade (National Association for the Education of Young Children, 1988; National Association of State Boards of Education, 1988; Shepard & Smith, 1988). If high-stakes pressure is already distorting instruction, what will happen if schools are evaluated in comparison to an inflated and escalating norm?

An obvious alternative, suggested by two of the original respondents to Cannell, is to develop new tests every year. Publishers could consider using the same equating procedures that allow several versions of the SAT and ACT to be used every year. Such proposals are apparently rejected out-of-hand because the costs are thought to be prohibitive. High school students pay $11.50 each to take the ACT. Counting the amortized cost of the initial purchase of booklets, districts pay $3.50 or more per pupil per year for off-the-shelf standardized tests. Obviously, states and districts would not be willing to maintain their current programs at three times the cost. But the more that the integrity of scores becomes an issue, the more states and districts might be willing to consider testing one-third as many grade levels

or different subjects every year. Other solutions include the use of sampling procedures (pupil sampling or matrix sampling) that reduce the incentive and the means to teach the test, or state developed tests such as writing assessments and student portfolios that make a greater effort to capture in the assessments the full extent of learning domains. States that have not been able to afford to develop their own tests might consider forming consortia to create tests that have more expansive content and procedural safeguards (such as multiple forms) to prevent teaching to test items.

References

Bracey, G.W. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. *Phi Delta Kappan, 68*, 683-686.

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

Cohen, S.A., & Forman, D.I. (1987). *Scoring high.* Westminster, MD: Random House School Division.

Darling-Hammond, L., & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85*, 315-336.

Drahozal, E.C., & Frisbie, D.A. (1988). Riverside comments on the Friends for Education report. *Educational Measurement: Issues and Practice, 7*(2), 12-16.

Jaeger, R.M. (1973). The national test-equating study in reading (The Anchor Test Study). *Measurement in Education, 4*, 1-8.

Koretz, D. (1986). *Trends in educational achievement.* Washington, DC: Congressional Budget Office.

Koretz, D. (1987). *Educational achievement: Explanations and implications of recent trends.* Washington, DC: Congressional Budget Office.

Koretz, D. (1988). Arriving in Lake Wobegon. Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8-15, 46-52.

Lenke, J.M., & Keene, J.M. (1988). A response to John J. Cannell. *Educational Measurement: Issues and Practice, 7*(2), 16-18.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1989). *Comparing state and district test results to national norms: Interpretations of scoring "above the national average"* (CSE Tech. Rep. 308). Los Angeles: UCLA Center for the Study of Evaluation.

Mehrens, W.A., & Kaminski, J. (1988, April). *Using commercial test preparation materials for improving standardized test scores: Fruitful, fruitless, or fraudulent?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

National Association for the Education of Young Children. (1988). *Position statement on standardized testing of young children 3 through 8 years of age.* Washington, DC: Author.

National Association of State Boards of Education. (1988). *Right from the start: The report of the NASBE task force on early childhood education.* Washington, DC: Author.

Phillips, G.W., & Finn, C.E. (1988). The Lake Wobegon effect: A skeleton in the testing closet? *Educational Measurement: Issues and Practice, 7*(2), 10-12.

Popham, W.J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*, 679-682.

Qualls-Payne, A.L. (1988). SRA response to Cannell's article. *Educational Measurement: Issues and Practice, 7*(2), 21-22.

Rhode Island Department of Education. (1988). *Testing Coordinator's Handbook, Rhode Island State Assessment Program, 1987-1988.* Providence: Author.

Shepard, L.A. (1988, April). *Should instruction be measurement-driven?* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Shepard, L.A., & Smith, M.L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. *The Elementary School Journal, 89,* 135-145.

Stonehill, R.M. (1988). Norm-referenced test gains may be real: A response to John Jacob Cannell. *Educational Measurement: Issues and Practice, 7*(2), 23-24.

Williams, P.L. (1988). The time-bound nature of norms: Understandings and misunderstandings. *Educational measurement: Issues and Practice, 7*(2), 18-21.

Woo, E. (1988, September 1). 40 grade schools cheated on skill tests, state finds. *Los Angeles Times,* p. 1.