DOCUMENT RESUME

ED 354 260 TM 019 520

AUTHOR Shepard, Lorrie A.
TITLE Chapter 1's Part in the Juggernaut of Standardized Testing.
REPORT NO TAC-B-291
PUB DATE Apr 92
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accountability; *Compensatory Education; Cost Effectiveness; Economically Disadvantaged; *Educational Assessment; Educationally Disadvantaged; Elementary Secondary Education; Evaluation Methods; *Federal Programs; Identification; Program Evaluation; Program Improvement; *Standardized Tests; Student Evaluation; *Testing Problems; Test Results; Test Use
IDENTIFIERS *Alternative Assessment; Education Consolidation Improvement Act Chapter 1; Hawkins Stafford Act 1988; High Stakes Tests; Performance Based Evaluation

ABSTRACT

The place for standardized testing in Chapter 1 evaluation is discussed. There is substantial evidence available on the negative effects of high-stakes standardized testing, and there is a clear link between Chapter 1 requirements and the amount of testing in most school districts. Standardized testing is usually used to identify eligible students, evaluate the Chapter 1 program, and hold individual schools accountable. It is argued that each of these purposes can be served better by other means. Alternative assessments are needed for Chapter 1 use, but any such assessments must be removed from the tyranny of normal curve equivalent gains. Any system that is devised should be subjected to its own cost-benefit evaluation to determine the costs and side effects of program improvement monitoring. Five overhead projection figures used in the presentation are included. (SLD)

CHAPTER 1's PART IN THE JUGGERNAUT OF STANDARDIZED TESTING

Lorrie A. Shepard
University of Colorado, Boulder

Chapter 1's Part in the Juggernaut of Standardized Testing

Lorrie A. Shepard

University of Colorado, Boulder

In my presentation this afternoon, I will make several key points in summary fashion. Most of my argument relies on evidence that has been thoroughly documented and discussed elsewhere. Therefore, I will supply you with references to those more complete analyses.

Research evidence on the negative effects of high-stakes standardized testing

Major findings from research on the effects of standardized testing are summarized in Overhead 1. First we know--from Linn et. al's more complete replication of the Cannell study (where all 50 states were above average) and from our own studies where students in high-stakes districts were retested with unfamiliar tests--that standardized tests scores can be corrupted as a valid indicator of student learning thus giving an inflated impression of achievement.

However, "teaching-to-the-test" practices which account for most of the distortion in test scores are of much greater concern because of what they do to instruction than for their effect on measurement validity. High-stakes tests narrow the curriculum. For example, elementary teachers

---

A symposium presentation at the annual meeting of the American Educational Research Association, San Francisco, April 23, 1992.

under pressure to raise reading and math scores report reducing or eliminating instruction in social studies and science. In addition, research in the last 5 years has found that high-stakes testing can distort even the way that basic skills are taught. Rather that practicing for norm-referenced tests for 2 or 3 weeks just before they are given, many teachers have altered their modes of instruction throughout the school year. Worksheets and classroom tests are designed in the exact image of standardized, multiple-choice tests creating a fill-in-the-blank mentality. Students are given more and more practice in recognizing one right answer rather than generating their own problems and solutions. As explained by the Resnicks (1992), teaching in the mold of standardized tests carries forward erroneous learning theory assumptions from the early part of this century. Practice on decontextualized rote-level skills denies students the opportunity to develop conceptual understanding and problem solving abilities.

The role of Chapter 1 requirements in creating a test-driven curriculum for both regular and Chapter students

The research cited above does not pertain specifically to Chapter 1 students or testing requirements. However, there is a clear link between Chapter 1 requirements and the amount of testing installed in most school districts. Many school districts find it easier to test all students on a schedule that will satisfy Chapter 1 demands than to pull Chapter 1 students for separate testing or to administer two

separate programs. In California, for example, the state assessment program provides better school and district accountability information than norm-referenced tests, yet most districts have created duplicate testing programs because the state assessment does not provide every-pupil, every grade data necessary for Chapter 1 evaluation. An extensive report just released by the Congressional Office of Technology Assessment (OTA, 1992), documents other ways in which Chapter 1 has added to the local testing burden. Note that the higher the proportion of Chapter 1 students in a district, the greater the incentive to make the district's testing program conform to Chapter 1 requirements. Thus there is likely to be more norm-referenced testing of all pupils in urban school systems.

Once the decision has been made to test all students in a grade on norm-referenced tests, it is a simple step to the negative instructional effects outlined in the first section. Once collected scores must be publicly reported; and our research on the nature of high-stakes pressure tells us that school rankings in the newspaper are the single most potent influence on test-driven instructional practices. This means that the decision to administer norm-referenced tests in every grade shapes the character of instruction for all students in the regular classroom.

Furthermore, the focus on NCE gains as the only coinage of school-improvement evaluations exaggerates the negative instructional effects of testing for Chapter 1 students.

Consider that teachers are coping with what they rightly perceive to be an irrational evaluation system. The lower performing their students are, the more likely it is that students' pretest performance was off the scale of the required grade-level test. This may also mean that real gains are off the scale and indistinguishable in NCE units from chance scores. It is small wonder then, with the increased stakes created by school improvement scrutiny, that teachers are likely to play it safe and teach in ways that closely resemble standardized test demands. For example, they could drill students on recognizing the main idea in short reading passages (consistent with standardized tests) rather than trying to have students connect story understandings to their own experiences (á practice that is pedagogically sound but that has not been amenable to standardization on traditional tests).

Chapter 1 pullout programs often mean students working in isolation on low-level worksheets, a finding that is well documented by the research of Dick Allington and others (Allington, 1991). Prior to the current reauthorization, the Whole School Day Study (Birman, et al., 1987) concluded:

> that Chapter 1 students may tend to have limited exposure to higher order academic skills. In that study's sample of schools, most Chapter 1 elementary reading and mathematics projects provided students with few opportunities to engage in higher order skills. In reading, for example,

> students were taught phonics and vocabulary and taught to read words or sentences. They were rarely asked to read paragraphs or stories or to construe meaning from text. In mathematics, students practiced computation skills and seldom applied mathematics facts to solving problems. At the secondary level, Chapter 1 classes offered a greater variety of instructional content, in part reflecting greater variation in achievement levels among high school students. More often than not, however, Chapter 1 reading and mathematics instruction in secondary schools also focused heavily on lower order skills. (p. 86)

Given the higher stakes now imposed by the fear of triggering school improvement status, there is no reason to believe that the incentives to teach to norm-referenced accountability measures have been reduced. The Congress was aware of the problem of low-level instruction and in the 1988 law specifically called for greater attention to student achievement in higher order analytical, reasoning, and problem-solving skills. However, the federal regulations maintain the demand for NCEs. Shifting ever so slightly to the "comprehension" and "problem-solving" subtests of norm-referenced tests is hardly sufficient to mitigate the negative effects of teaching to the test or to honor the intention of Congress in any meaningful way.

In Overhead 3, I list three primary purposes for the existing testing requirements: to identify students eligible for services, to evaluate the Chapter 1 program, and to hold individual schools accountable. This list is admittedly an oversimplification as evidenced by the more complete list of uses provided by the Office of Technology Assessment report (1992) and shown in Overhead 4. Nonetheless, the three purposes I have identified constitute the major uses for testing and subsume nearly all of those listed by OTA.

It is my contention that each of purposes for which standardized tests are currently used can be served better by other means. Because of the negative effects of the current system on instruction, Congress and the Department are obliged to consider the feasibility of these alternatives.

First, tests are unnecessary for valid and accurate identification of those students most in need of Chapter 1 services. From very old measurement studies comparing teacher ratings and test scores as well as more recent studies on special education referrals, I argue that teachers are not very good at making normative comparisons to children in other schools. However, they are quite accurate in ranking the relative skill levels of children within their own classrooms. Therefore, if a quota system were used that established the number of children to be served on socio-economic grounds, teachers could be relied

upon to refer the children most in need of services. Given that some local discretion is already permissible, it is likely that very nearly the same population would be served with or without test selection rules.

Second, a national evaluation of Chapter 1 could be conducted much more thoroughly and rigorously using a national probability sample rather than every pupil testing. For example, the outcome measures used in a national study could be much broader using a matrix sampling approach like National Assessment and would not be taught to in the same way as traditional norm-referenced tests. Furthermore, if only a sample were being studied it would be possible to give more valid pretests and come to a more accurate understanding of gains for students who are functioning in the lowest segment of the NCE scale. My recommendation in this regard is similar to that suggested by the Office of Technology Assessment:

> Congress could obtain national data on Chapter 1 through a well-constructed, periodic testing of Chapter 1 children, similar to the way NAEP is used to assess the progress of all students. This assessment would rely on sampling (rather than testing of every student) and could be administered less frequently than the current tests. In addition to relieving the testing burden on individual students and reducing the time devoted to testing by teachers, principals,

and other school personnel, this procedure could also result in higher quality data. (p. 35)

I would go further than the OTA report in one respect. As has been seen in the past, it does no good for there to be flexibility in what is permitted at the federal level if state-level requirements build in rigidity. Therefore, it is essential that the 1993 reauthorization consider an overhaul of the entire system, especially the school improvement provisions considered next. Otherwise national evaluation will simply be added on top of the existing system.

Finally, I consider alternatives to the current program improvement guidelines for holding local schools accountable for student progress. Although my suggestions in this arena are tentative and pose numerous technical and logistical difficulties, the wisdom of pursuing alternatives must be judged in light of the serious inadequacies of the present school improvement model. Those who think that the present model only needs fine tuning, will obviously be unwilling to tackle the problems that shifting to a new model would entail.

The limitations of the current model have been articulated previously. Slavin and Madden (1991) illustrated how the current accountability guidelines may discourage early interventions, and reward both teaching to the test and grade retentions. Similarly Clayton (1991) described the misdirection of effort and discouragement that

can occur if schoolwide projects are falsely judged to be unsuccessful. Studies by Bushner (1991) and Davis (1991) suggest that NCE gains at the school level are fraught with error. To these complaints I would add my own concerns that administering norm-referenced tests to low-achieving students serves no instructional purpose, and to the extent that children are functioning below the range of the test provides an insensitive measure of what they gain from the program.

The alternatives I propose are summarized in Overhead 5. If data from individual schools do not have to be in all the same metric to serve the purposes of national evaluation, then it is possible to think more carefully about what kind of data could be collected locally to hold schools accountable. Teachers could be asked to collect data consistent with good instructional practice. This might consist of informal reading inventories, performance assessments, or graded reading and math materials. By graded materials I mean things like a series of stories and pieces of literature that have been selected to mark a continuum of reading difficulty. Although there would be metric problems to be solved if one wanted to demand comparability between districts and schools, it would be possible for each locality to decide on a different way of measuring progress.

In my view, once the national evaluation question has been solved by a separate study, site-to-site comparability

is far less important than ensuring the instructional relevance of testing done to every single Chapter 1 child. However, for those who are concerned about the trustworthiness of locally reported data, it would be possible to check on the integrity of apparent gains either by a system of auditing visits or by selective retesting of grades or schools. From a different perspective, it might be more worthwhile to ignore checking on gain scores entirely and instead use audits to verify that teachers have an adequate understanding of how their students are functioning and are delivering instruction appropriately.

There is a great deal of talk about the development of alternative assessments for use in Chapter 1. I concur that such measures are needed. However, it is unlikely that most local schools can develop assessments of sufficient rigor and meet current guidelines for equating to norm-referenced tests. Furthermore, however rich assessments are for instructional purposes they will prove to be insensitive evaluation measures for the state and district, if they must be transformed by statistical equating to a narrow band on the within-grade NCE scale. Therefore, it is essential that school level evaluations be removed from the tyranny of NCE gains. I recommend that funds currently spent on Technical Assistance Centers to support the machinery of norm-referenced testing and NCE aggregation be redirected toward developing more instructionally relevant measures with particular attention to new reporting metrics.

In conclusion, whatever system is devised should be subjected to its own cost-benefit evaluation. What are the costs and side-effects of program improvement monitoring? Are the claimed benefits supported empirically? For example, aggregate NCE gains for Chapter 1 students were remarkably stable for the 10 years prior to Hawkins-Stafford legislation. Are there shifts in more recent data that can reasonably be attributed to massive program improvement efforts, or have the national data remained unperturbed?

References

Allington, R.L. (1991). Children who find learning to read difficult: School responses to diversity. In E.H. Hiebert (Ed.), Literacy for a diverse society: Perspectives, practices, and policies. New York: Teachers College Press.

Birman, B.F., Orlan, M.E., Jung, R.K., Anson, R.J., Garcia, Moore, M.T., Funkhouser, J.E., Morrison, D.R., Turnbull, B.J., & Reisner, E.R. (1987). The current operation of the Chapter 1 program. Washington, DC: U.S. Department of Education.

Bushner, D.E. (1991, April). Fall-to-fall testing versus spring-to-spring testing: What is the impact on a local community's Chapter 1 evaluation? Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Clayton, C.E. (1991). Chapter 1 evaluation: progress, problems, and possibilities. Educational Evaluation and Policy Analysis, 13, 345-352.

Davis, A. (1991). Upping the stakes: Using gain scores to judge local program effectiveness in Chapter 1. Educational Evaluation and Policy Analysis, 13, 380-388.

Kober, N. (1991). The role and impact of Chapter 1, ESES, evaluation and assessment practices. Washington, DC: Office of Technology Assessment. PB 92-127646.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), Changing assessments:

Alternative views of aptitude, achievement and instruction. Boston: Kluwer Academic Publishers.

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

Shepard, L.A. (1991). Will national tests improve student learning? Phi Delta Kappan, 232-238.

Slavin, R.E., & Madden, N.A. (1991). Modifying Chapter 1 program improvement guidelines to reward appropriate practices. Educational Evaluation and Policy Analysis, 13, 369-379.

U.S. Congress, Office of Technology Assessment. (1992). Testing in American Schools: Asking the Right Questions, OTA-SET-519. Washington, DC: U.S. Government Printing Office.

Overhead 1

Research evidence on the negative effects of high-stakes standardized testing

1. When test results are given high-stakes by political pressure and media attention, scores can be corrupted, thus giving a false impression of student achievement.

2. High-stakes tests narrow the curriculum. Tested content is taught to the exclusion of non-tested content.

3. High-stakes testing misdirects instruction even for the basic skills.

4. The kind of drill-and-practice instruction that tests reinforce is based on outmoded learning theory, what the Resnicks refer to as the decomposability and decontextualization assumptions. Rather than improve learning, such instruction actually denies students opportunities to develop thinking and problem-solving skills.

Overhead 2

The role of Chapter 1 requirements in creating a test-driven curriculum for both regular and Chapter students

Chapter 1 school-improvement requirements, which effectively mandate annual reporting of NCE gains, tie local districts to every-pupil, every-grade administration of norm-referenced tests.

Thus Chapter 1 drives local testing programs for regular education; and

The focus on NCE gains as the only coinage of school improvement exaggerates the negative instructional effects of testing for Chapter 1 students.

Turning to the "comprehension" and "problem-solving" subtests of norm-referenced tests is hardly sufficient to mitigate negative effects or to honor Congress' intention that attention be focused on higher-order reasoning skills.

Overhead 3

Purposes for testing. Are there alternatives?

Student eligibility

Program evaluation

School accountability

Uses of standardized tests in Chapter 1

LEA uses:

- identifying which children are eligible for Chapter 1 services and establishing a "cutoff score" to determine which children will actually be served;
- assessing the broad educational needs of Chapter 1 children in the school;
- determining the base level of achievement of individual Chapter 1 children before receiving services (the "pretest");
- assessing the level of achievement of Chapter 1 children after receiving services (the "posttest"), in order to calculate the change data required for national evaluations;
- deciding whether schools with high proportions of low-achieving children should be selected for projects over schools with high poverty;
- establishing goals for schoolwide projects;
- determining whether schoolwide projects can be continued beyond their initial 3-year project period;
- annually reviewing the effectiveness of Chapter 1 programs at the school level for purposes of program improvement;
- deciding which schools must modify their programs under the "program improvement" requirements;
- determining when a school no longer needs program improvement;
- identifying which individual students have been in the program for more than 2 years without making sufficient progress; and
- assessing the individual program needs of students that have participated for more than 2 years.

Uses by Congress and the Department of Education:

- national evaluation of Chapter 1;
- justifying continued appropriations and authorizations;
- weighing major policy changes in the program;
- targeting States and districts for Federal monitoring and audits; and
- contributing to congressionally mandated studies of the program.

(Office of Technology Assessment, 1992)

Overhead 5

Purposes for testing. Are there alternatives?

Student eligibility

SES quotas followed by teacher nomination of individual students

Program evaluation

National probability sample, in-depth assessment on a periodic cycle

School accountability

Demands for data collection should be consistent with good instructional practice.

Chapter 1 teachers should keep records charting progress (using informal inventories, performance assessments, graded reading and math materials). Aggregate gain scores should be reported from classroom assessments.

State and district requirements for aggregate data should not reinvent the every-pupil, every-grade NRT model but should check on the integrity of gains reported from the classroom. For example, local records could be subject to audit or retesting by randomly selecting grades or schools.

Funds currently spent on Technical Assistance Centers should be redirected toward developing more instructionally relevant measures of student progress with attention to reporting metrics.