DOCUMENT RESUME

ED 375 149 TM 022 097

AUTHOR D'Agostino, Jerome
TITLE Improving the Identification of Schools for Chapter 1 Program Improvement.
PUB DATE Apr 94
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Compensatory Education; Disadvantaged Youth; Educational Assessment; Educationally Disadvantaged; Elementary Secondary Education; *Evaluation Methods; *Identification; Item Response Theory; Models; *Norms; Program Evaluation; *Program Improvement; Regression (Statistics); *Standards; Statistical Bias; Student Evaluation; Urban Schools; Validity
IDENTIFIERS *Education Consolidation Improvement Act Chapter 1; Iowa Tests of Basic Skills

ABSTRACT

Technical problems with norm-referenced achievement testing that can lead to the erroneous evaluation of schools for Chapter 1 Program improvement is discussed, and an alternative testing model is presented. The history of Chapter 1 testing and evaluation policies is briefly reviewed, and problems with the norm-referenced model are explored. Data from a large urban district with an extensive Chapter 1 system for the Iowa Test of Basic Skills are used to demonstrate the way in which regression bias can lead to unintended consequences that impair the validity of an identification scheme. An alternative testing model, based on item response theory, is proposed that would allow schools to develop their own assessment devices. To avoid inequities, the ideal testing system would require schools receiving Chapter 1 funds to administer two tests to their students, a locally developed assessment and the state test. The Federal government could establish criteria for program identification that the school would meet, such as 75% of students attaining the school's standards, and the state test would ensure that the school's standards were appropriate. (Contains 19 references.) (SLD)

Improving the Identification of Schools for Chapter 1 Program Improvement

Jerome D'Agostino

University of Chicago

Paper presented at the 1994 meeting of the American Educational Research Association, April, New Orleans, LA.

Since the mid-1970s, norm-referenced achievement testing has served as the main indicator of Chapter 1 program success. Although it was originally developed to gauge the overall effectiveness of the program, the norm-referenced model is now used to identify schools in need of program improvement. Many researchers have questioned the model's suitability for fulfilling this additional function, yet most of these researchers have emphasized the technical deficiencies of the model or its inability to provide schools pertinent information about student learning. Few of them have investigated how these technical problems can lead to the spurious identification of schools for Chapter 1 Program Improvement.

This paper has four main purposes. First, I will briefly explain the history of Chapter 1 testing and evaluation policies. Then I will provide an overview of the problems plaguing the norm-referenced model. It will become evident that most research in this area has not studied how these problems can misclassify programs for Chapter 1 Program Improvement. I will demonstrate with data from a large urban district that one limitation of the model, regression bias, can lead to unintended consequences that greatly impairs the validity of the identification scheme. Finally, I will propose an alternative testing model that would be based on Item Response Theory (IRT) and that would allow schools to develop their own assessment devices.

A Brief History of Chapter 1 Evaluation

Chapter 1 evaluation has primarily consisted of testing student achievement. In fact, in Chapter 1 research, the terms "testing" and "evaluation" are often used interchangeably. Evaluation has been part of Chapter 1 from the program's beginning, but its purposes and procedures have changed with each reauthorization.

The original law required local school districts to provide a yearly report on program effectiveness, but it did not stipulate specific evaluation procedures. In 1967, revisions were made that required the Office of Education to report each year to Congress on the effectiveness of the program. A system was developed where each LEA orchestrated their own evaluation, and then submitted the results to the states. The states then summarized the LEAs' reports, and the Office of Education compiled the states' summaries in a final document. Apparently, the quality of the local evaluations varied dramatically (Davis, 1991).

Congress improved evaluation procedures in the 1974 reauthorization by mandating that an objective testing system be developed. RMC Research was contracted to design the evaluation model. The company created a system that allowed LEAs to choose one of five approved plans. Due to difficulties in implementing some of the models, the number of options was reduced to three, which were a norm-group comparison, a control-group comparison, and a regression design (Horst, Tallmadge, and Wood, 1975). The norm-referenced model soon became the preferred design because it was more convenient and required less technical expertise to orchestrate than the other two models.

The 1974 reauthorization not only altered Chapter 1 evaluation procedures but it also changed its purpose as well. Up to that time, testing mainly served as a means of appraising the success of the program at the national level; it was not intended to offer schools pertinent information they could use to improve their Chapter 1 programs. The 1974 amendments modified the role of evaluation by stating that, "one of its purposes was to offer LEAs a way of assessing their programs and to serve as a tool for program revision and improvement" (U.S. House of Representatives, 1974, p. 20). Although Congress wanted evaluation to fulfill this new role, they did not modify testing policies. Consequently, schools were expected to improve their programs partially based on test results that offered them limited knowledge about the progress of their students. This point will be revisited in the next section.

In 1988, the new role for Chapter 1 evaluation was firmly established by the Hawkins-Stafford reauthorization. The law required schools receiving funds that did not demonstrate "substantial yearly progress" to develop a Program Improvement Plan (PIP). If a school was placed in improvement status, they had to submit a plan to their LEA, and if the school could not demonstrate progress after two years, it was to enter into a joint plan with the educational agency of the state. Program progress was to be measured by the norm-referenced model that had been intact since the mid-1970s. If a school's Chapter 1 students had a mean gain (in NCE units aggregated across grades) of zero or less on the approved standardized tests, or on tests equated to those tests, the school's program was to be identified as in need of improvement (U.S. Department of Education, 1989, p. 21774).

The Hawkins-Stafford amendment defined these rules as the minimum standard a state could set to identify schools. The law strongly encouraged states to set more stringent NCE standards, and it allowed them to use additional tests other than the approved ones. Since the first year of the reauthorization, 15 states have set higher NCE standards (e.g., 75 percent of students at a school must show gain) and some states have employed their statewide test as another identification measure (Gittleman, 1992).

The 1988 law, however, did not permit LEAs or SEAs to supplant the federally certified tests with innovative assessment tools. In fact, the LEA of Jefferson County, Kentucky had asked the U.S. Department of Education if they could exclusively use their state performance-based assessment system for identifying schools, but they were denied permission (Gittleman, 1992).

Problems With The Norm-Referenced Model

The program improvement model created in 1988 is still in use today even though many educational researchers have seriously questioned its adequacy. Before describing these problems, it is important to understand some of its key assumptions. Instead of comparing Chapter 1 students' test scores with a traditional comparison group (as is the case with the other two RMC designs), the model bases comparison on national test norms. A student's standing relative to the norm group on the pretest is compared to the student's standing relative to the norm group on the posttest. It is assumed that without program services, the

student would stand at the same level on the posttest as the pretest, which would be indicated by a zero change score (Tallmadge & Wood, 1976). This is commonly referred to as the equipercentile model.

Most criticisms leveled against the present model fall into one of two categories; concerns related to measurement issues or concerns related to its negative consequences in shaping the Chapter 1 curriculum.

Measurement Issues

The technical merit of using simple gains from norm-referenced tests as indicators of program quality has been seriously questioned by many researchers. Jaeger (1979) demonstrated that differences between two schools' mean NCE gains varied depending on the standardized test used for comparison. In other words, two schools may have different mean NCE gains simply because they administered different tests. Another study by Anderson (1991) showed that the same percentile change from pretest to posttest translated to different NCE gains across various tests.

A few years following the development of the norm-referenced model, other researchers, notably Linn (1979), began to criticize the logic of the equipercentile assumption. In order for the assumption to be sound, norming samples would need to be nearly identical to Chapter 1 groups. Because students are usually selected for the program based on low test scores, they are often not representative of norm groups. Consequently, the groups used to norm tests often do not serve as adequate control groups necessary to assess treatment related growth. If the groups are discrepant from Chapter 1 samples, the norm-referenced model may underestimate treatment effects. Further, Linn (1979) questioned the soundness of assuming that norm groups serve as "no-treatment" controls given that many students constituting these groups receive other types of special services and may come from more privileged school districts.

If Linn's argument is accurate, then the higher a Chapter 1 student scores on a pretest, the more difficult it is for that student to make a gain on the posttest. For example, as pretest scores increase, comparison students become more proficient, and thus, more difficult to surpass in terms of growth. Converting test scores into NCE units is supposed to yield interval scales, but if growth becomes more difficult for participating students with higher pretest scores, it is questionable if the norm-referenced model can render scales with equal intervals (Linn, 1981; Anderson, 1991).

The regression to the mean effect, where students with extreme scores on a pretest tend to drift toward the average score on the posttest independent of any intervention, appears to be a prevalent problem in Chapter 1 testing. Any model that employs simple gains as measures of student growth assumes that initial status is not correlated with gains. Because this is often not the case in Chapter 1 testing, Davis (1991) was able to demonstrate that schools with above average mean pretests were significantly more likely to be categorized in

program improvement than schools with means below average.

In order to demonstrate another unintended repercussion of the regression bias when identifying schools in need of improvement, I analyzed data consisting of Iowa Test of Basic Skills reading 1991 (pretest) and 1992 (posttest) NCE scores of 21,987 Chapter 1 students from a large urban school district with an extensive Chapter 1 system. Similar to the findings of Davis (1991), I discovered that students' pretest scores were significantly and inversely related to simple gains, $r = -46$, $p < .001$. Further, a significantly smaller percentage of schools with above median pretest scores had positive mean gains than schools with mean pretests below the median, X^2 (1, $N = 132$) = 10.56, $p < .001$. In other words, schools with larger proportions of students who had higher pretest scores were more likely to be categorized in Program Improvement compared to schools with less students who scored relatively high on the pretest.

In the district in which the data were gathered, schools were required to select students for Chapter 1 services that score below the third stanine (NCE = 34) on the ITBS. Schools, however, may also have chosen students based on other criteria such as teacher recommendations. From other sources of information, I found that many schools in this district were more likely to select students in lower grades than upper grades by using these other criteria. Consequently, schools may have unintentionally increased their risk of being placed in Program Improvement by selecting students for Chapter 1 services based on indicators other than test scores, because many students selected in this manner may have scored above third stanine on the pretest.

Further statistical analyses of the data indicated that this effect was apparently occurring. Thirty-eight percent of the second graders in the sample had pretest scores above the third stanine compared to twenty percent of the sixth graders. The correlation between the percentage of second graders at each school was negatively related to school mean gain, -37, $p < .01$, and the correlation between the percentage of third graders in a program was also negatively related to mean gains, -.18, $p < .01$. None of the correlations between percentages of fourth through eight graders and mean gains were significant.

Categorical analyses indicated that schools with higher concentrations of second graders in Chapter 1 were more likely to be identified as being in Program Improvement than schools with lower concentrations. Schools were classified as above or below the median percent of second grade students served in Chapter 1. The schools above the median percentage of second graders were more likely to be classified in Program Improvement (i.e., had a mean gain of 0 or less) than schools below the median percentage of second graders, X^2 (1, $N = 132$) = 5.94, $p = .01$. The same analysis was performed using schools' percentages of third graders served. The same trend was evident, but the relationship was not significant. For seventh graders, the reverse relationship was discovered. Schools with above median percentages of seventh graders in Chapter 1 were more likely to have positive mean gains than schools under the median, X^2 (1, $N = 132$) = 4.13, $p < .05$. Thus, if schools used criteria other than pretests to identify students for Chapter 1, they could have inadvertently

increased the likelihood of being targeted for Program Improvement.

The study I performed along with Davis' (1991) findings clearly reveal the inadequacies of the norm-referenced model for accurately and fairly identifying schools in need of programmatic change. Perhaps the most egregious result of employing an imprecise testing model for Chapter 1 has been how school personnel have responded to being identified for improvement. The recent National Implementation Study showed that if school staff felt the testing system yielded inaccurate results, they were less likely to seriously commit time and effort to improving their program (Abt Associates, 1992).

Curriculum Issues

Apart from the measurement problems of the norm-referenced model, Chapter 1 testing procedures may contribute to a narrowing of the curriculum and adoption of a belief that students learn best by being passive recipients of knowledge. Because program success is based on mean reading and math gains for each school, many Chapter 1 teachers feel compelled to spend most of their instructional time on these two subjects. Other topics such as science and social studies are frequently presented at the end of the school day if time permits, and very few Chapter 1 dollars are spent on instructional approaches that are not reading or math related.

Like any other "high-stakes" testing situation, Chapter 1 instruction is often focused on teaching students the skills they will need to answer the types of items commonly found on standardized multiple-choice tests. Often enough, students are asked to complete boring worksheets that contain items with one correct answer and emphasize basic skill mastery (Shepard, 1992).

Perhaps the most vital concern of Chapter 1 educators relates to the usefulness of the information provided by norm-referenced test score results for designing an effective program improvement strategy. Many school personnel have complained that reports of student gains on norm-referenced tests provide little pertinent knowledge about what students actually learned. Schools need to know what their students can do and the areas where they need help if they are to design a viable PIP.

As mentioned in the prior section, when program improvement was incorporated with testing in the 1988 reauthorization, testing procedures were not properly realigned to offer schools necessary information about the specific types of skills their students had developed. To place the evolution of evaluation within Peterson, Rabe, and Wong's (1992) model of program maturity, present testing practices were established when Chapter 1 administrators were concerned with imposing and enforcing regulations, while program improvement was institutionalized later when the federal emphasis had shifted to local flexibility. The current testing program was developed to assure compliance rather than to give schools constructive feedback on the educational progress of students.

Now that Chapter 1 has evolved to the point where stakeholders are attending more to improving programs, the number of people involved in the program that feel the norm-referenced model is outdated and unsuited for the new emphasis is steadily increasing. As Gittleman (1992) aptly states, "The greatest flaw with the program improvement process is the federally required system of measuring educational progress using norm-referenced standardized tests," (p 2). Considering all of the problems with the present system, it may not only be an inaccurate means of identifying schools in need of improvement, it may serve as a hinderance to schools desiring to change.

The Future of Chapter 1 Testing

Testing policies will likely change when Congress reauthorizes the program this year. The Clinton Administration has proposed to Congress an amendment that would allow states to use their assessment systems to satisfy federal accountability requirements (U. S. Department of Education, 1993). Presently, the House is drafting their version of the reauthorization. and their bill (HR 6) would modified the Clinton Administration's proposal to grant states the flexibility to develop their own accountability systems by requiring states to do so in order to receive Chapter 1 funds. The Senate has yet to write a reauthorization bill, but from floor debates over Goals 2000, it appears the Senate will agree to retain the Clinton proposal but reject the more prescriptive House plan. Many Republicans in the House also oppose the requirement stipulated in HR 6, so it is likely that the Clinton plan to grant states flexibility to use or develop their own tests, rather than require them to do so, will be part of the final law. Given the widespread disfavor over the norm-referenced model, most states will probably capitalize on the new law and adopt their own testing system.

If states are granted this option, the goals and properties of Chapter 1 evaluation will probably vary from one state to the next. Indeed, four states (Nevada, New Hampshire, Wyoming, and Iowa) have yet to develop state testing systems ("Student-Assessment," 1994). According to comprehensive data collected on state assessments by the North Central Region Education Lab (Perlman, personal communication, April, 1994), both test types and their purposes clearly differ across states. For instance, some states use them for school accountability while others use them for graduating students. Vermont, Maine, Kentucky, Arizona, New Mexico, and California use some form of a student portfolio. Other states such as Alabama and Arkansas have norm-referenced basic skills tests that are not unlike the presently approved tests used for Chapter 1 evaluation. Still other states use either criterion-referenced tests, writing samples, performance-based instruments, or a combination of these formats. Also, some states test in every grade while others test in a portion of grades, and some test every student while others do not.

Although most Chapter 1 reformers agree that the norm-referenced model does not provide the type of information schools need to develop a good PIP, research examining the impact of the newly developed state assessments on schools' program improvement progress is virtually nonexistent. Research is lacking in this area because federal requirements do not allow the state assessments to be the central form of program monitoring and accountability.

Consequently, even though reformers have been advising Congress to abolish the present testing regulations, it is still quite speculative that the new state tests will rectify all the inherent problems of the norm-referenced model and provide schools the valuable data they need to draft effective improvement plans.

Furthermore, states have not embraced the same theory of educational reform (Darling-Hammond, 1993). Some states have developed a system predicated on the assumption that more central control is needed, and that schools must work harder toward reaching externally set standards. Other states attempt to build teacher capacity, emphasize more decentralized control, and employ testing systems that foster local school development. The type of assessment system created by a state reflects the theoretical approach to reform the state has adopted.

Although it can be argued that Chapter 1 Program Improvement was initially an outgrowth of the reform theory espousing more centralized control, it is quite clear that federal policies have begun to mirror the tenets undergirding the reform theory based on developing individual school capacity. Thus, testing systems more aligned with the latter reform version will probably be more useful for schools participating in the Chapter 1 Program Improvement of the future.

Darling-Hammond's (1993) research has demonstrated that allowing schools to create their own alternative assessments is a powerful way to foster school-level reform initiatives. She claims, "There are ripple effects throughout the entire school organization when teachers begin to ask questions such as these: What do we want students to be able to do? What can we develop as a means for evaluating their knowledge and abilities?," (p. 760). Given that the primary intention of Program Improvement will be to encourage schools to grapple with curricular and assessment issues, tests that can be developed locally would appear to enhance these internal school processes.

A Proposed Testing Model

I would like to propose a testing system that could be implemented on a wide-scale and serve as an improvement over the present model by addressing both the measurement and curricular issues I have discussed. As stated, Congress intends to loosen federal requirements in an effort to foster more school-level decision making. A testing system that would probably be most aligned with these efforts would be one that allows schools (or groups of schools) to develop or employ their own assessment devices. Of course, this approach could easily create an inequitable accountability system because schools' tests would likely vary in difficulty. Thus, an ideal testing model would provide schools some degree of latitude in designing their assessments while simultaneously retaining an acceptable level of standardization across schools, districts, and states.

On way of accomplishing this ideal testing system would be to require schools receiving Chapter 1 funds to administer two tests to their students; a locally developed

assessment and the state test. Teachers would be asked to set their own standards which would be manifested in their assessments. The federal government could establish a general criteria for Program Improvement identification, such as 75 percent of participating students at a school must attain the school's standards. The state test would serve to ensure that schools do not set unusually low or high standards. For instance, if a school claimed that its students met their standards, but the state test revealed that students were below the state expectations, the school would be asked to revise its test. This procedure would be another form of Program Improvement, but now the focus of improvement would be on the school's curriculum instead of student outcomes.

In order to solve the technical problems created by the norm-referenced model, the school tests could be scored using IRT procedures. IRT is based on the properties of objective measurement that Wright and Stone (1979) refer to as "person-free item calibration" and "item-free person measurement." That is, people's estimated abilities should not depend on the items they took and the difficulty of the items should not depend on the group of people who took them. IRT advocates argue that latent trait scaling is a more accurate means of creating interval scales than norm-referenced techniques. Measuring student growth is now commonly performed by using IRT methods because they conveniently solve the measurement problems created by norm-referenced testing that I have described.

This testing system could easily be integrated within Chapter 1 Program Improvement. Part of drafting a PIP would entail developing the school assessment system. This would encourage schools to reflect on their academic goals and foster teacher collaboration and reflection. Instead of being overly concerned about teaching their students test-taking skills every spring, teachers would need to work together to develop a strategy for increasing students' progress over multiple school years. Schools could work together to develop a local item bank, which may be more feasible given the likely difficulty schools would experience while attempting to generate their own assessments within time and resource limitations (see Wright, 1977, for uses of IRT measurement). Wilson (1991) has made slightly similar recommendations for Chapter 1 testing.

Although Program Improvement became a major reason for Chapter 1 testing in 1988, the present norm-referenced model does not provide schools the type of information related to their students' progress necessary to make effective program alterations. Given the maturity of Chapter 1 as a federal program, it is time to create an evaluation system that is a true service to schools rather than as a simple means of program compliance. Allowing schools to develop their own assessment systems within the templates of state standards seems to be one potentially effective way of getting school personnel to invest valuable time in redirecting their Chapter 1 programs.

References

Abt Associates. (1992). The Chapter 1 Implementation Study: Interim report. Cambridge MA.

Anderson, J. (1991, April). Using the norm-referenced model to evaluate Chapter 1. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Darling-Hammond, L. (1993). Reframing the school reform agenda. Phi Delta Kappan, 73, 753-761.

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.

Gittleman, M. (1992). Chapter 1 program improvement and innovation across the states. An overview and state profiles. Washington, DC: Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED 350 379).

Horst, D., Tallmadge, G., & Wood, C. (1975). A practical guide to measuring student achievement. Number 1 in a series of monographs on evaluation in education, U.S. Department of Education.

Jaeger, R. (1979). The effect of test selection on Title 1 project impact. Educational Evaluation and Policy Analysis, 1, 33-40.

Linn, R. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), Education evaluation methodology (pp. 84-109), Baltimore, MD: The Johns Hopkins University Press.

Linn, R. (1979). Validity of inferences based on the proposed Title 1 evaluation models. Educational Evaluation and Policy Analysis, 1, 23-32.

Peterson, P., Rabe, B., & Wong, K. (1991). The maturation of redistributive programs. In A. Odden (Ed.), Education policy implementation (pp. 65-80). Albany, NY: SUNY Press.

Shepard, L. (1992, April). Chapter 1's part in the juggernaut of standardized testing. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Student-Assessment mandates now in 46 states, survey finds. (1994, February 2). Education Week, p. 4.

Tallmadge, G. & Wood, C. (1976). User's guide: ESEA Title 1 evaluation and reporting

system. Mountain View, CA: RMC Research Corp.

U. S. Department of Education. (1993). Improving America's Schools Act of 1993. Washington, DC: U.S. Department of Education.

U. S. Department of Education. (1989). Final regulatory guidance for the implementation of Chapter 1 of P.L. 100-297. Federal Register, 21774-21794.

U. S. House of Representatives. (1974). Elementary and Secondary Education Amendment of 1974 (House Report 93-805, Serial Set 13061-1). Washington, DC: Government Printing Office.

Wilson, M. (1992). Educational leverage from a political necessity: Implications of new perspectives on student assessment for Chapter 1 evaluation. Educational Evaluation and Policy Analysis, 14, 123-144.

Wright, B. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. & Stone M. (1979). Best test design. Chicago, IL: MESA Press.