DOCUMENT RESUME

ED 350 315 TM 019 011

AUTHOR	Anderson, Judith I.
TITLE	Using the Norm-Referenced Model To Evaluate Chapter 1.
PUB DATE	Apr 91
NOTE	20p.; Notes for a Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1992).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Compensatory Education; Elementary Secondary Education; Equated Scores; *Evaluation Methods; *Federal Programs; *Models; National Norms; *Norm Referenced Tests; Pretests Posttests; Program Effectiveness; *Program Evaluation; School Districts; Test Norms
IDENTIFIERS	Aggregation (Data); Education Consolidation Improvement Act Chapter 1; Hawkins Stafford Act 1988

ABSTRACT

In response to growing frustration over the lack of information about the national effectiveness of the Chapter 1 program, Congress enacted the Education Amendments of 1974. Section 151 of the Amendments directed the U.S. Office of Education to develop evaluation models that would allow school district data to be aggregated to provide national estimates of program effectiveness. The norm-referenced model was the most easily applied of the alternatives developed. This model substitutes test norms for a traditional comparison group. Posttest standing relative to the norm group is compared with pretest standing relative to the norm group. The 1975 document, "A Practical Guide to Measuring Project Impact on Student Achievement," specified the conditions in which the norm-referenced model could be used. Several difficulties have arisen in implementing the models, but school districts today are still required to evaluate their Chapter 1 projects. Requirements enacted in 1988 mean that districts essentially must use nationally normed tests or tests equated to nationally normed tests to measure student achievement in both basic and more advanced skills. Test norms and norm-referenced tests are reviewed, with attention to measurement error, the effects of high-stakes testing, and the relevance of national norms as a comparison group. Ways in which information on program effectiveness could be better provided are discussed. Five figures and one table illustrate the discussion. (SLD)

Using the Norm-Referenced Model to Evaluate Chapter 1

Judith I. Anderson

Notes for a Symposium Presentation at the Annual Meeting of the American Educational Research Association

April 5, 1991

NOTE: This paper is intended to promote the exchange of ideas among researchers and policy makers. The views are those of the author, and no official support by the U.S. Department of Education is intended or should be inferred.

Development of the Title I (Chapter 1) Evaluation and Reporting System

Title I of the Elementary and Secondary Education Act of 1965 was enacted to provide special educational assistance to students in areas impacted by poverty. The law mandated that local school districts report annually on program effectiveness, but did not require specific reporting formats or standards. Subsequent attempts to estimate the effectiveness of the program nationwide were less than overwhelmingly successful. (See Hope Associates, 1979, for a review of the early history of Title I. Title I is now Chapter 1 of the Education Consolidation and Improvement Act, or ECIA.) In response to growing frustration over the lack of information on the national effectiveness of the program, Congress enacted the Education Amendments of 1974. Section 151 of the Amendments directed the U.S. Office of Education (USOE) to develop evaluation models which would allow district data to be aggregated to provide national estimates of program effectiveness.

USOE responded by contracting with the RMC Research Corporation to develop evaluation models which would provide a national estimate of the effectiveness of Title I, and, in 1975, USOE released "A Practical Guide to Measuring Project Impact on Student Achievement" (Horst, Tallmadge, and Wood, 1975), which provided TItle I projects with assessment methods which allowed evaluation results to be aggregated across projects nationwide. The "Practical Guide" provided five evaluation models for districts to consider when measuring project impact:

- A posttest comparison using matched groups;
- Analysis of covariance;
- Special regression (either regression projection or regression discontinuity models);
- A general regression model; and
- The norm-referenced model.

All models were designed with the goal of providing "as clear and unambiguous an answer as possible" to the question "How much more did pupils learn by participating in the project than they would have learned without it?" (Horst, Tallmadge, and Wood, 1975, p. 1)

Not all of the five models in the "Practical Guide" turned out to be practical for a majority of school districts, and by 1976 the five models had been reduced to three: a comparison group model, a special regression model, and the norm-referenced model. Each of these models existed in two forms, one a norm-referenced version; the other a non-normed version. (See Tallmadge and Wood, 1976.) There were still problems. Most school personnel in 1976 lacked either the expertise or the computers to easily implement the regression model, and the comparison group model was more often than not impossible to implement due to program rules which specified that the most needy schools and students were to receive services. This left, by default, the norm-referenced model.

The Requirements

The norm-referenced model substitutes test norms for a traditional comparison group. Students' posttest standing relative to the norm group is compared with their pretest standing relative to the norm group—essentially, it is assumed that in the absence of a treatment percentile standing remains the same, and, given that a treatment had been provided, any change in percentile standing can be attributed to the treatment. This assumption is commonly referred to as the "equi-percentile" assumption. The "Practical Guide" gave as strengths of the model that:

> "Where no comparison group is available, the norm group provides a plausible estimate of no-treatment posttest scores. Even where a comparison group is available, unless it comes from the same population as the treatment group, the Norm-Referenced Model offers a more defensible estimate of posttest performance at substantially less cost and effort than a comparison-group model."

The weakness was that:

> "The validity of the model rests on the assumption that the achievement status of a particular subgroup remains constant *relative to the norm group* over the pre- to posttest interval if no special treatment is provided. Empirical support for this assumption is minimal. It is conceivable that some subgroups would move up and others move down in the normal course of events. When the norm group is like the treatment group, the plausibility of the underlying assumption is greatly enhanced; thus, for example, norms for gifted children would be best for assessing a project serving such pupils." (p. 72 of Horst, et. al. 1975)

However, the guidelines for model selection gave the clear impression that the norm-referenced was a model of last resort, with a decision tree that led one to the model only after the other two were rejected as unfeasible. Some comfort was provided to the poor soul left with no choice but to use the norm-referenced, as it was noted that "Despite the clear order of preference, all three of the models should yield valid evaluation results if properly implemented. Feasibility and other practical considerations may well counterbalance scientific rigor, and it is certainly the case that a well implemented Model A [the norm-referenced model] will yield more credible results than a poorly implemented Model B [the comparison group model] or C [the special regression model]." (Tallmadge and Wood, 1976, page 21.)

In the 1975 "Practical Guide", users of the norm-referenced model were informed that they must use standardized tests, use the same level of the test for both pre- and posttesting, select their program participants on a measure other than the pretest, and conduct both pre- and posttesting on the norming dates used by the test publisher. A t-test was used to compare students' actual average posttest score (using standard scores, since normal curve equivalents, or NCEs had not yet been added to the system) with the expected posttest score (based on the pretest mean), and the resulting t value was tested for statistical significance at the .05 level. If the difference was statistically significant, one next checked for educational significance, with a general rule of thumb being to determine whether the observed posttest scores exceeded the expected posttest scores by a third of a standard deviation.

Users were allowed to use tests without national norms (e.g., district or state tests) if the tests were equated to nationally normed tests. The inclusion of an option for non-normed tests was perhaps practically and politically necessary: while the movement against standardized testing was less well-defined 15 years ago than today, critics existed. The typical alternative of the day, however, was the district or state criterion-referenced test, and in order to use these tests, districts were required to equate them to tests with national norms. For the most part, this proved difficult, either because of the costs involved in the additional testing required or because the tests measured different types of skills and equating was impossible. After all, if the locally developed test was so similar to the nationally normed test that there was a very high correlation between the scores on the two tests, the odds were that the tests were similar enough that it was easier to just use the standardized test, and most districts did just that.

Districts were allowed to assess student achievement using either a fall-to-spring testing cycle or an annual cycle (e.g., spring to spring or fall to fall), and they were cautioned to:

(1) use the same level and form of the test at pretest and posttest,
(2) use functional level testing,
(3) test within 2 weeks of the publisher's norming dates, and
(4) select students for the program on some other measure than their pretest scores.

In the 1976 guide, raw scores were converted to NCEs, and testing for statistical significance of the gains was eliminated. By 1981, the revised "Evaluator's Reference" proclaimed that "All NCE gains greater than zero are good!"

After the models were developed, USOE began working with State and local education agencies to implement them, and funded 10 Technical Assistance Centers (TACs) to help school districts with their Chapter 1 evaluations. The 1979-80 school year was the first year of full district implementation of the models.

The Reaction

Complaints about the models began immediately. A survey by Hope Associates (1979) found complaints from State education agencies, Technical Assistance Centers, and some USOE staff that evaluations were not useful for school districts, that the requirements excluded "process" evaluation, and that the Title I "evaluation" system was really a Title I "reporting" system. Congressional aides, on the other hand, felt that they were developing a system that would be useful for SEAs and LEAs.

USOE staff at the time—of which I was one—noted in response that they were required to develop a system that would provide a means of aggregating district information to provide a *national* estimate of program effect, and that districts were not precluded from supplementing the basic system with more information. In addition, USOE routinely challenged critics to propose alternative systems which would meet the basic requirement of providing a means of aggregating data to provide national estimates of program effectiveness. While there was interest in making the evaluations be as useful as possible for school district personnel, there was a definite feeling among many of the evaluators involved that the data would have limited utility for very small projects because of the measurement error associated with the model and the appropriateness of comparison to national norms. While many of these errors would balance out when the data were combined at the national level, school personnel were advised to base program decisions on a variety of sources of evidence and not to make judgments about either pupils or programs based on one test score, or, for that matter, on any other single piece of evidence.

By 1981, Model A was enough a part of the evaluation landscape that Linn (1981) included in a textbook discussion on measuring pretest-posttest performance changes. He noted that with Model A "[t]here is a certain intuitive appeal to the assumption that in the absence of special interventions students will tend to maintain the same relative standing in achievement. Intuitive appeal is not a very adequate basis for an evaluation assumption, however. Furthermore, even if the assumption were justified under idealized conditions, it still would lead to biased estimates of project effects as it is implemented in practice." He also noted that "[t]he model rests on a strong assumption for which there is no adequate basis"—i.e., the equi-percentile assumption—and provided examples of both factors that could lead to positive bias (e.g., regression effects) and factors which could lead to negative bias, including dissimilarity of the project group and the group on which the test was normed: "By their very nature, special programs are often designed for groups that are quire dissimilar to the usual norming sample. Due to this dissimilarity, the expectations based on the norms may be too high or too low." Linn concluded (p. 95) that "[i]n summary, the constant NCE approach to defining normal growth and thereby providing a means for project evaluation lacks an adequate justification. In some applications, it may be seriously defective and result in biased estimates of the effects of an educational program."

Other authors reanalyzed existing data bases to determine the actual growth of compensatory education students over time and, thus, the validity of the equi-percentile assumption. Kaskowitz and Norwood (1977) produced a USOE sponsored review of utility of the norm-referenced model for the evaluating Project Information Packages (PIPs). (PIPs were specific projects which were being disseminated by USOE for possible adoption by other sites.) They noted that evidence from the first year PIP evaluation indicated that compensatory education students may be the students who lose ground over time, relative to the norm population, and they analyzed Metropolitan Achievement Test (MAT) data for several different groups of students to attempt to determine whether that was indeed the case. They noted that the MAT "empirical growth curves do indicate that the equipercentile growth curve tends to underestimate expected posttest scores for extremely low pretest scores and tends to overestimate expected posttest scores for extremely high pretest scores across grades levels and tests." When they examined actual performance for a group of students who were not in the Follow-Through program (i.e., were not receiving compensatory education), they found an average loss of approximately 6 percentile points each year from the spring of the first grade to the spring of the third grade in reading; in math, there was a 6 point drop from the first to the second grade, and no further drop at the end of third grade. There were difference patterns of score changes for white and minority group students, but the direction of changes varied across the grades examined. The authors concluded that "[u]se of the norms based on the standardization group will lead to an expected posttest score that will be too high for students ordinarily in compensatory programs, especially minority students who have pretest scores that are not extremely low" and "[f]or pupils with extremely low pretest scores, the equal percentile assumption leads to a predicted standard score that is much lower than what was observed." (Kaskowitz and Norwood, 1977, p.55) In other words, depending on which students you have, your expected posttest score can be either too high or too low.

Tallmadge (1982) used Sustaining Effects Study (SES) and California Achievement Test (CAT) national norming data to assess the norm-referenced evaluation methodology. He provided project level (i.e., school within grade) data from both the SES and CAT, as shown in Table 1, which provided evidence that in the aggregate – that is, across many projects – the norm-referenced design provided reasonable estimates of program impact. The SES data provided information on schools which did not have Title I or other compensatory education programs, while the CAT data base in all likelihood included some students who were receiving either Title I or other compensatory services. An examination of the data from both sources shows that, on average, the equi-percentile assumption is more or less valid. The CAT data, which most likely include remedial students, show slight positive biases (which may not be biases at all, but actual program effects) and the SES data, which do not include compensatory education students, show overall gains near zero, but with grade differences, particularly at Grade 2, where students without compensatory services actually lost ground relative to the norm.

As can be seen from Table1, there is considerable variation at the project level, with a large percentage of the projects experiencing a loss in percentile standing from pretest to posttest. Thus, while overall – that is, at the national level – the equipercentile assumption appears to more or less hold, at the individual project level, this is not the case. This would be expected, of course: all measurement contains error, some positive, some negative. Tallmadge noted that "Larger projects would produce smaller standard deviations but, as a rough approximation, sample sizes of at least 20 students will be required before real treatment effects of 4 or 5 NCEs can be reliably discriminated from random errors." (p. 110) Tallmadge also noted that "When implemented at the project level, the treatment that the norm-referenced model evaluates is the students' total school experience. If the intervention itself is only part of the total school experience, as is the case with Title I projects, it is at least theoretically impossible to separate the impact of the intervention from that of the test of the school program. If a superior school produced above-average growth rates, the gains made by its Title I students would exceed expectations even if the Title I project itself were ineffective. Similarly, if a below-average school were responsible for abnormally low growth rates, and effective Title I project would be made to appear less effective."

A minor skirmish appeared in the literature when Powers, Slaughter, and Helmick (1983) reported, based on the analysis of actual 7th and 9th grade students' test scores that there was "a pattern of overestimation of gains" which was "evidence of the inappropriateness of the equipercentile assumption." Tallmadge (1985) responded by reinterpreting Powers et. al.' s data to reach an opposite conclusion and presented claims that the national norms are "quite robust" and provide "acceptably accurate" no-treatment expectations for low achieving students in large, medium, and small LEAs of varying size and urbanicity.

Table 1

Estimates of Project Effectiveness Using the Norm-Referenced Model: Project Data Presented by Tallmadge, 1982

	Number of Projects	Mean NCE	Standard Deviation
Sustaining Effect Study Data Base			
Grade 2	103	-.95	5.36
Grade 4	99	.48	3.89
Grade 6	97	.19	3.14
California Achievement Test Data Base			
Grade 2	102	1.79	5.56
Grade 4	109	.80	3.53
Grade 6	109	1.49	3.85

For the most part, however, once it was clear that the system was not going to go away, State and local education agency (LEA) personnel implemented the models without much more ado, and the Department of Education began producing annual reports on Chapter 1 participation and effectiveness which used the data submitted by the States. Certain patterns in the data were noted. First, as had been noted in some prior studies, the estimates of student achievement based on fall-to-spring gains were considerably higher than those based on annual testing, and they did not seem to be sustained over time. That is, if one tested the same students the following fall, the estimate of their achievement gain was considerably lower than the estimate based on the spring posttest score. The losses were variously attributed to summer forgetting, inaccurate norms, stake-holder bias, and other causes. Whatever the cause, however, the gains obtained with fall-to-spring testing were not a good measure of actual student achievement over a full year. Second, gains tended to be higher in the lower grades than in the higher grades. Whether this was due to differences in test norms, to needier students being served in the higher grades (who might be more difficult to help), or to the greater effectiveness of intervention at the lower grades was never thoroughly investigated. Third, mathematics projects tended to show higher gains than reading projects. Again, this may have been due to differences in norms, differences in the types of students served (the math students tended to have higher pretest percentiles than the reading students), or real differences in project effectiveness for reading and mathematics. The practical effect of these differences was that schools and districts which concentrated services in the lower grades and in mathematics projects, and which tested fall-to-spring, tended to look more successful that did schools which had reading projects, served students in the higher grades, and which used annual testing. For this reason, achievement data were kept separate by grade level, testing cycle, and subject matter.

Requirements Today

School districts today still are required to use the models to evaluate their projects. In fact, use of the models has been expanded beyond the original intent of providing national estimates of program effectiveness to use for program improvement.

Section 1019 of Public Law 100-297, which was enacted in 1988, mandates that local education agencies (LEAs) must conduct evaluations at least once every three years in accordance with the national standards developed under Section 1435. Essentially, this means that they must use nationally normed tests, or tests equated to nationally normed tests, to measure student achievement. The law now requires them to measure student achievement in both basic and more advanced skills. The "advanced skills" measurement in practical application means that in reading they must give a reading comprehension subtest and in math, a math problems and applications subtest, and report on these separately. The use of the evaluation models has been potentially expanded by Section 1021, which deals with school program improvement. Any school which "does not show substantial progress" towards meeting its program goals or "shows no improvement or a decline in aggregate performance of children served under this chapter for one school year as assess by measures developed pursuant to section 1019" must develop a school program improvement plan which includes technical assistance, alternative curriculum, improving coordination with the regular classroom program, evaluates parent involvement, and provides inservice training.

The regulations (Federal Register, 1989) and Policy Manual (1990) provide further guidance. LEAs must evaluate their programs at least once every three years, and, for their reading, mathematics, and language arts programs in grades 2 and above, they must use norm-referenced tests (or tests equated to nationally normed tests) to provide an estimate of what their achievement would have been without the Chapter 1 program, give the pretest and posttest at least 12 months apart, and calculate the estimate of gain using the NCE metric. Certain positive changes in requirements are to be found. One is that the fall-to-spring evaluation cycle has been eliminated. Projects must use either a spring-to-spring or a fall-to-fall testing cycle to measure project impact. The Policy Manual also emphasizes that LEAs must conduct evaluations based on both aggregated student achievement *and* other desired outcomes: for example, success in the regular classroom and attaining grade level proficiency.

All schools that serve 10 or more students in their Chapter 1 programs (across grade levels and subjects) must develop a school improvement plan. As part of the plan, achievement data are aggregated by subject area for grades 2 and above in each school building. (Programs at the pre-kindergarten, kindergarten, and grade 1 level are exempted from using normed achievement tests for evaluation due to concerns about test reliability and validity for children at those levels.) Each school must look at its pretest to posttest change in achievement standing, and "No gain or a decline in aggregate performance scores in the subject that is the primary focus of the Chapter 1 program, as measured according to the national standards for evaluation, causes a school to be identified for program improvement." (Policy Manual, page 154) Schools are not allowed to place confidence bands around the estimates of performance to determine whether the gains or losses are statistically significant. Testing for statistical significance is considered to be "an effort to avoid program improvement", not an attempt to determine the meaningfulness of the scores. Schools which believe that a few extreme scores may have influenced their gains may use the median rather than the mean – but only if all other schools in the district also use the median – and schools worried about measurement error are encouraged to use multiple measures to assess individual student progress.

In the 1987-88 school year, over one million Chapter 1 reading students and over 625,000 mathematics students in Grades 2 through 12 had both pre- and posttest scores. These students represented (some States sample school districts) over 1.6 million reading students and nearly 975,000 math students – a large number, but less than two-thirds of the reading and math students at these grades (Sinclair and Guttman, 1990). The estimated percentages of Chapter 1 students for whom achievement data were provided ranged from 74 percent for Grade 3 reading to only 14 percent for Grade 12 reading.

A Brief Review of Test Norms and Norm-Referenced Tests

In order to understand what confidence one can place on standardized test scores and the interpretation of norm-referenced test results, it helps to keep in mind several conditions which influence test scores, including the effect of measurement error, the effects of "high stakes" testing, and the relevance of the norm group as a comparison for the students being assessed.

Measurement Error

Most people realize that students' test scores are just estimates of their performance and that the scores will vary from testing to testing due to chance. However, not everyone realizes just how much scores are likely to vary. How likely is it that a student's score might vary by several raw score points by chance alone? It is very likely. The standard error of measurement on the ITBS Level 9 Form J for grade 3 students is 2.8 raw score points for the reading comprehension subtest and 1.94 raw score points for the math problems subtest. For the MAT-6, Elementary Level, Form L, grade 3 students have a standard error of measurement of 3.1 raw score points on the reading comprehension subtest and 2.2 on the math problem solving subtest. Thus, about one-third of the time, a student's "true" score could be at least 3 points higher or lower on the MAT and ITBS reading comprehension subtests and about 2 points higher or lower on both math problems subtests. Obviously, if a project has large numbers of students, measurement errors of this sort balance out. However, for projects with very few students, this will not necessarily be the case, and observed gains or losses may be due as much to measurement error as to program quality (or lack thereof.)

Furthermore, many people do not realize what difference a few items can make in a student's percentile rank on a norm-referenced test, although many authors have commented on this. Shepard (1989) examined both the California Achievement Test (CAT) and the Stanford Achievement Test (SAT) and determined that a gain of one raw score point at the median in reading, language, and mathematics could translate into a percentile gain of from 2 to 4 points. I examined the Iowa Tests of Basic Skills (ITBS) and the Metropolitan Achievement Test (MAT) Reading Comprehension and Mathematics Problems subtests appropriate for students in grade 3, and found a similar situation, although the actual gain varies by test and percentile standing. It is obvious that one raw score point must translate into more than one percentile point gain when you look at the number of items on the tests: given that most subtests have fewer than 99 items, a gain of more than one percentile point for each raw score point is inevitable. What is less clear is what effect this may have at different points on the score distribution.

On the ITBS Form J Reading Comprehension subtest, which has 42 items, one additional item correct for a grade 3 student assessed in the spring results in a percentile change of from 0 to 5. (See Figure 1.) The NCE gains range from 0 to 9. A student going from 12 to 13 items correct would move from the 9th to the 11th percentile (the 22nd to the 24th NCE), but a student going from 13 to 14 items correct would move from the 11th to the

15th percentile (the 24th to the 28th NCE.) On the MAT-6 Form L Reading Comprehension, a change of one raw score point can make a difference of as much as 4 percentile points and 6 NCEs, or – for students at the very lowest levels – it can make no difference at all. (See Figure 1)

On the ITBS Level 9, Form J, Math Problems subtest, which has only 24 items, one additional item correct buys a grade 3 student tested in the spring between 0 (for those students in the chance range) to 10 percentile points. (See Figure 2.) A student answering 12 items correctly would obtain a percentile rank of 34, but one answering 13 items correctly would obtain a percentile rank of 41! On the MAT Elementary Level, Form L there are 30 items, and each additional item correct can be "worth" up to 7 percentile points. A student answering 17 items correctly has a percentile score of 43; one extra item correct moves the percentile score to 50.

Figure 1
Effect of Raw Score Changes on Percentile Standing – Reading Comprehension

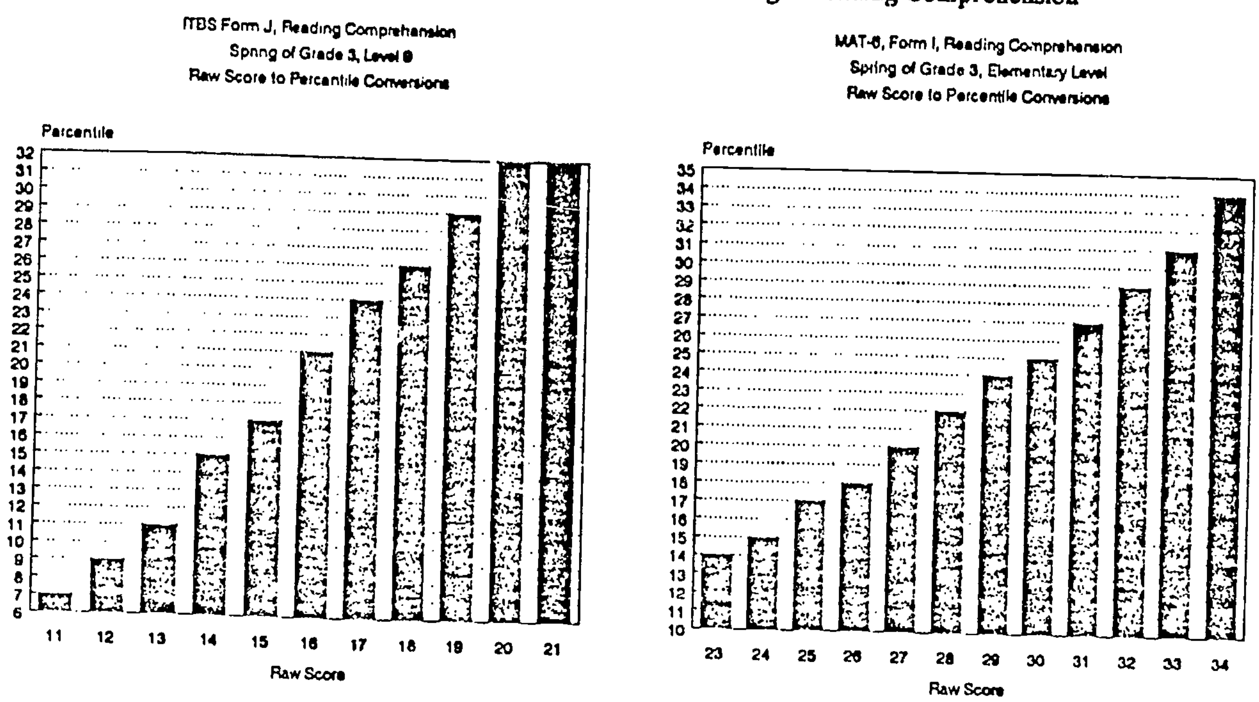
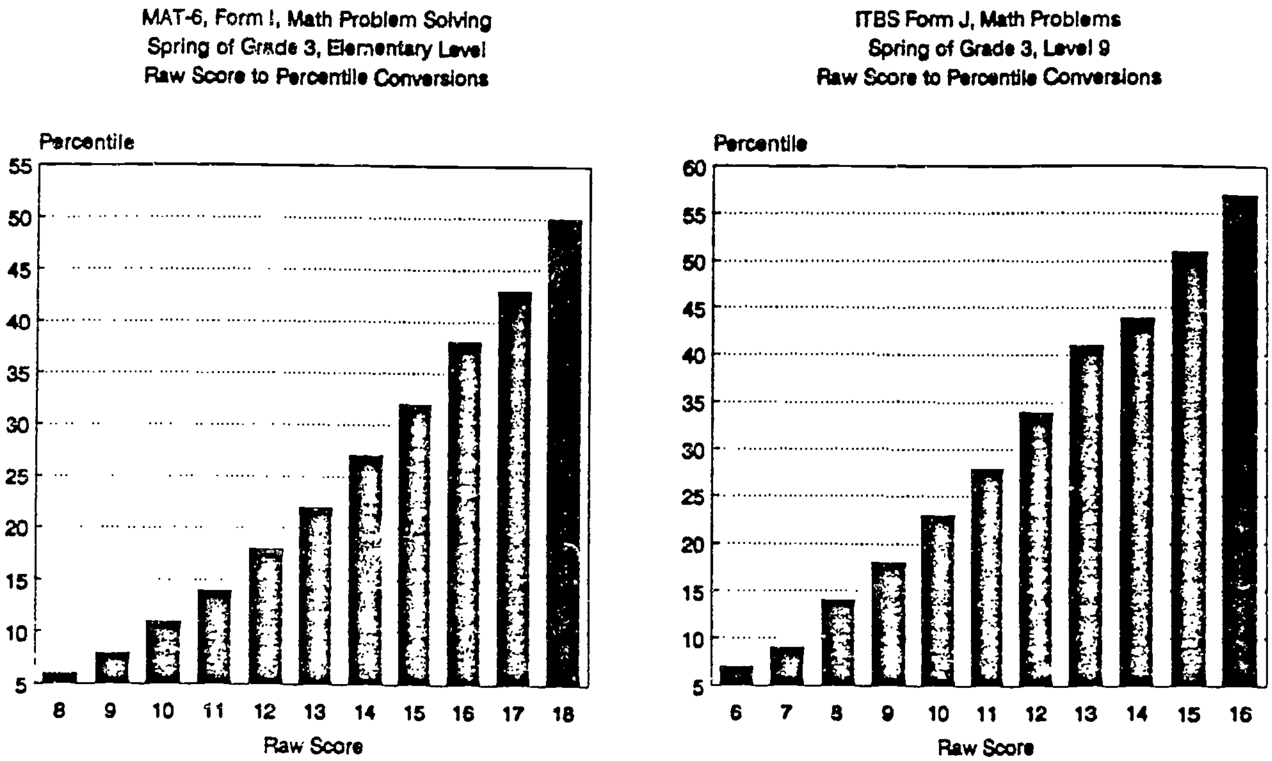


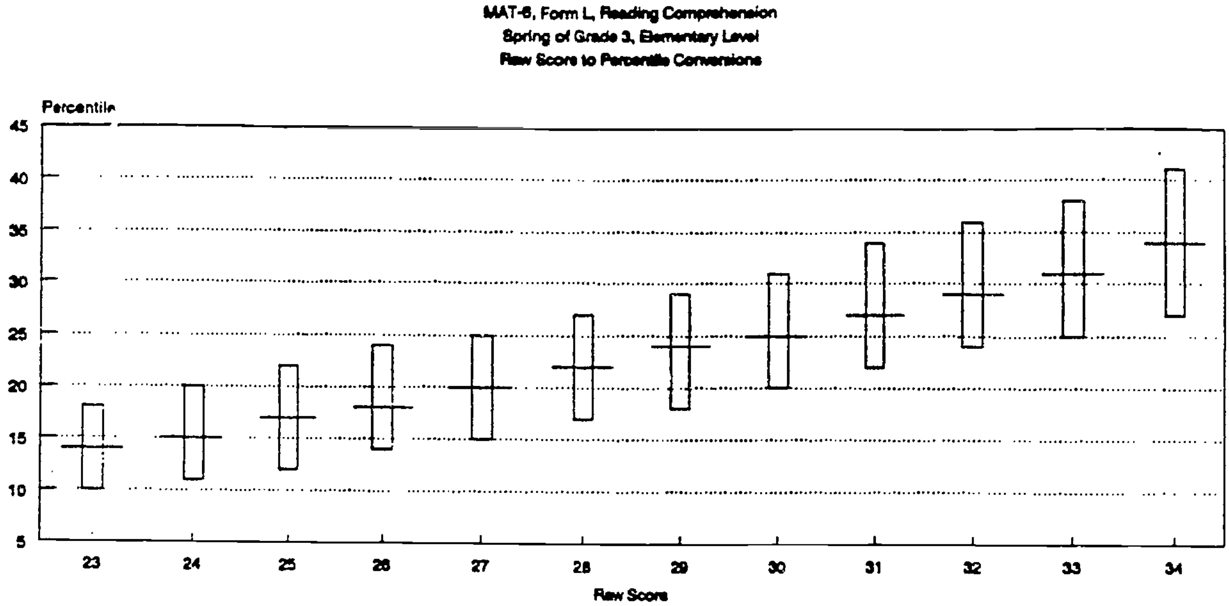
Figure 2
Effect of Raw Score Changes on Percentile Standing—Math Problem Solving



The effect of measurement error on a student's percentile standing can be seen in Figure 3. As noted above, the standard error of measurement on the MAT–6, Form L, Reading Comprehension subtest for grade 3 students is 3.1 raw score points. For a grade 3 student who obtains a raw score of 27 on the MAT reading comprehension, there is a one-third chance that her "true" percentile standing is less than 15 or greater than 25—a wide range, indeed.

The practical effect for projects which are relying on just one score to assess achievement is that instruction can appear to have helped—or to have hurt—students when the observed change is due not to the instruction but to error. As the number of students increases, the errors will balance out, but for small projects, there can be large observed changes from chance alone.

Figure 3
Confidence Bands for Individual Student Scores:
Bands are for Plus or Minus One Standard Error of Measurement



Effect of "High Stakes" Testing

Measurement error, of course, is not the only factor which can effect students' scores in ways that can effect interpretation of project impact when using standardized tests. There have been numerous complaints about the effects of "high stakes" testing (see Herman, Golan, and Dreyfus, 1990, for one review.) Herman et. al. note that in the past, tests were not expected to alter curriculum and instruction, but with the advent of "high stakes" testing – e.g., situations where teachers, schools, and districts are rated or ranked based on achievement test scores – teachers are more likely to focus their instruction on those areas measured by the test. While some might find this positive – after all, if the test measures important areas, it makes sense to teach those areas – one should recognize that tests generally measure a fairly narrow set of objectives. In a survey of 85 teachers, Herman et. al. found that low socio-economic elementary schools give the most attention to test results, and the reaction is to engage in additional test preparation activities. They found that elementary school teachers spend the equivalent of several weeks on test taking strategies, with teachers with decreasing test scores engaging more often in these activities. Furthermore, there was a definite Chapter 1 effect: schools with larger numbers of Chapter 1 students were more likely to feel pressure to raise test scores and to spend more time not only trying to cover all of the curriculum but also on test preparation.

14

Shepard (1989) investigated possible reasons for inflated test score gains, many of her findings are elevant to problems which may occur in Chapter 1 program evaluation. She noted that:

> "Tests are selected to achieve the best possible match between the test and the curriculum. ... Once the test is chosen that best fits the curriculum, the pratical curriculum is adjusted further in response to the test."

As noted above, this is not necessarily bad – assuming that one selects a test to measure one's goals, focusing instruction towards those goals, and therefore to the test, makes sense. But, Shepard notes that:

> "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 were apparantly learning the tested content and other things as well when they took the unannounced test."

In addition, children in grade 1 through 3, and sometimes grade 4 as well, take a practice test first, unlike children in the norm sample, which further restricts one's ability to interpret the students' normative standing. Shepard concludes by stating that:

> "In this study we have been concerned primarily with what test-curriculum alignment and teaching to 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."

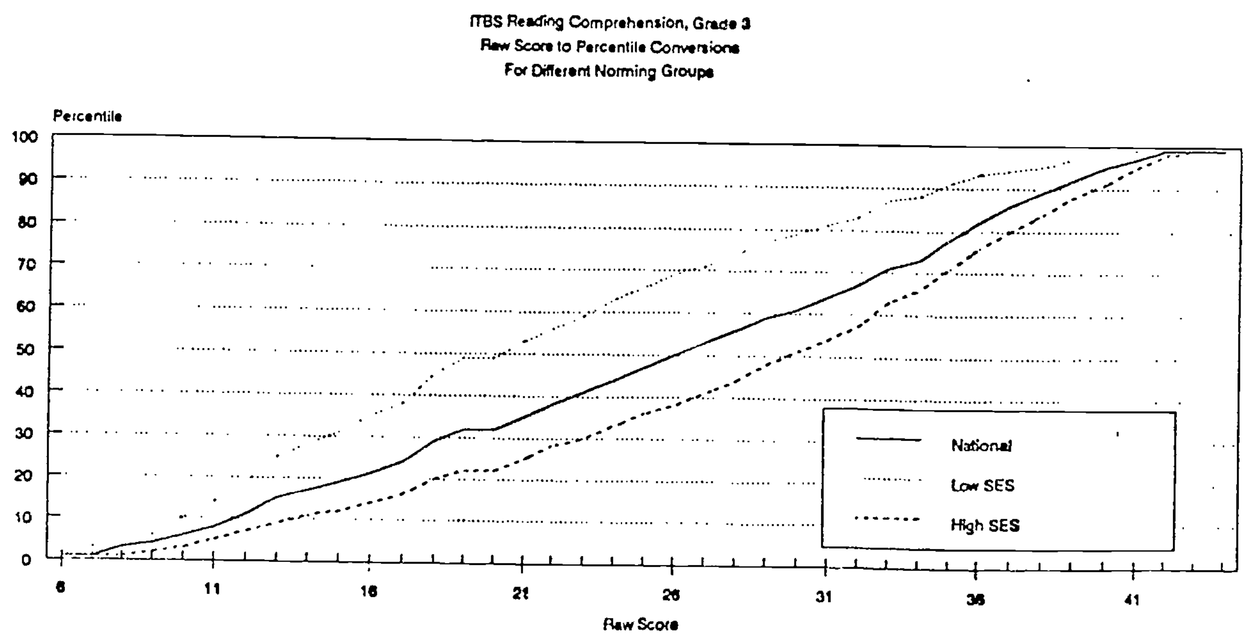
Given the difference that a few raw score points can make in students' test scores, and the pressure to produce project test score gains, it seems likely that Chapter 1 teachers would be tempted to focus their instruction on the narrow areas measured by the tests. In addition, inadvertent focus on specific vocabulary or topics which the teachers know are covered on the particular test that they are using can easily result in test score improvements which invalidate normative comarisons.

The practical effect is that teachers who do *not* narrow their curriculum to the areas covered by the test are at a disadvantage when compared to those teachers who do. If project – and teacher – success is measured by test score gains, more teachers are going to feel pressure to "teach to the test" in order to improve their students' scores. Unfortunately, these score improvements will not necessarily mean that the students are receiving improved instruction and curriculum.

Relevance of National Norms as a Comparison Group

The underlying rationale of the norm-referenced model is that the national norms provide a valid comparison group. If we are comparing projects nationwide with the national norms – the original intent of the model – it may. However, for individual projects, there may be less reason to assume that such comparison is valid. There is an implicit assumption that the principal factor in assuring that the comparison group is valid is the pretest score. However, the students at, say, the 15th percentile in the national norm group come from a variety of schools and communities. Some of the students who score at the 15th percentile are in schools which have very high concentrations of poor children, limited resources, and poor regular programs. Others are not. If the original premise of Chapter 1 is correct – that is, that schools with high concentrations of poor children have special problems – then the Chapter 1 teachers in these schools are being held to a more difficult standard than are the Chapter 1 teachers in less poor schools. At the reverse end of the scale, many Chapter 1 programs are in schools and districts which have a very low percentage of poor children. The Chapter 1 projects in these districts may have an "easier" comparison. The "Practical Guide" (1975) noted that one evaluation hazard to be avoided was the "use of non-comparable treatment and comparison groups" and noted that project personnel should look not just at pretest standing but also at differences in age, sex, race, or socio-economic status. While test publishers today are paying much more attention to ensuring that tests are free of sex and race biases, there are still different achievement patterns for students in high and low poverty areas. How much difference can this make? One clue comes from looking at the norms for schools in low and high socio-economic areas compared with the national figures on the ITBS Reading Comprehension subtest, which has 1985 norms for all three groups. As shown in Figure 4, the percentile equivalents for different raw scores differs

Figure 4
Differences Between Low Socio-economic Status and High Socio-economic Status Schools on the Iowa Test of Basic Skills



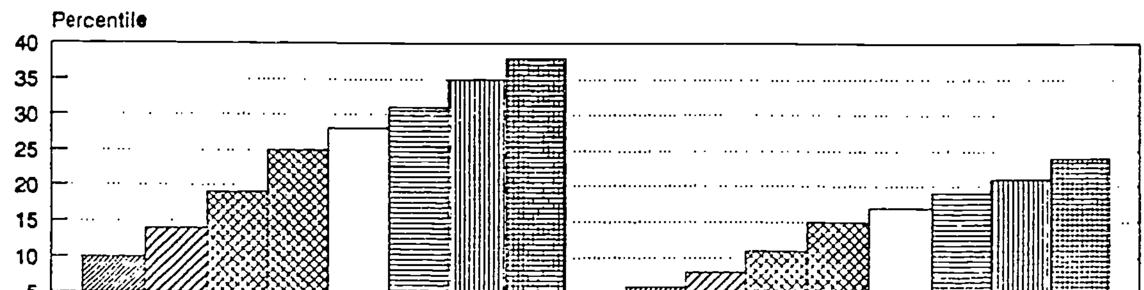
considerably, which is not a surprising finding given what is known about the relationship of school poverty and achievment.

The actual impact of raw score point changes on percentile changes may be less apparant, and is illustrated in Figure 5. At the low end of the percentile range, each additional raw score point results in a larger percentile change on the low SES norms than on the national norms. For example, for a student to move from the 19th to the 25th percentile on the national norms takes slightly over 2 additional raw score points (from 15 to 17+). However, only one additional raw score score point on the low SES norms (from 12 to 13) raises a student from the 19th to the 25th percentile on the low SES norms.

While is is essential that we maintain as a program goal that students in areas which are heavily impacted by poverty be provided with instruction and other support which enables them to catch up with children from other areas, we need to be equally aware of the difficulties that their teachers face, and measure the teachers' effectiveness against reasonable comparisons.

One additional problem that occurs today is that many students in the national norming samples for tests are actually receiving Chapter 1 or equivalent instruction, particularly at the lower grade levels. Nationwide, about 20 percent of children in grades 1 through 3 receive Chapter 1 services, although not all of them receive reading or math. To the extent that the norms include students served by remedial programs, comparison with the norms is not comparison with a no-treatment expectation but rather with an alternative treatment.

Figure 5
ITBS 1985 Norms for Schools in Low SES Areas and for the Nation
Grade 3, 1985 Spring Norms



Raw Score	Low SES Percentile	National Percentile
10	10	6
11	14	8
12	19	11
13	25	15
14	28	17
15	31	19
16	35	21
17	38	24

Raw Score to Percentile Conversion

Conclusions

The Chapter 1 (formerly, Title I) evaluation and reporting system was developed to provide a national estimate of the effectiveness of the Chapter 1 program — that is, to provide an estimate of how much more students learned with the program than they would have learned without it. The system is now used not only to provide an estimate of national effectiveness, but also to estimate effectiveness at the project level. It is time to consider revising this system, and to provide alternatives for measuring both project impact and national effectiveness.

Information on the national program effectiveness of Chapter 1 could be better provided by national samples and special purpose studies which collected more detailed information on samples of students rather than by forcing nearly all districts to test their students every year. The results of the current system provide limited information for small projects, and despite the large numbers of students tested — over one million in reading and over 625,000 in mathematics during the 1987–88 school year — these students represent less than two-thirds of the students in the program at Grades 2 through 12, and therefore the results may not even be a good representation of national program effectiveness. Furthermore, there is a danger that "high-stakes" testing may be causing teachers to narrow their curriculum to the limited areas covered by the test, as well as inadvertently teach to specific areas of the test. Both of these procedures invalidate comparison with the national norms, and the former may mean that students are provided with worse, not better, instruction.

Until alternatives are implemented, we need to take steps to minimize misinterpretation of the results of Chapter 1 evaluations. These include re-instituting the requirement to test changes in scores for statistical significance and to add confidence bands to our estimates of individual project effectivenss. This is especially important for small projects. We also must make it clearer that no decisions on Chapter 1 projects should be made based on only one piece of data, and work with project staff to provide meaningful alternatives, including assisting them with designing evaluation systems which assess student changes over longer periods of time, provide for alternative measurements of student progress, and add assessments of the instructional program which provide teachers with valid information about how to modify their programs in ways which will actually improve student achievement.

References

Herman, Joan, Golan, Shari, and Dreyfus, Jeanne. Paper presented at the annual meeting of the California Educational Research Association, Santa Barbara, California, November 1990.

Hope Associates, *Performance Review of USOE's ESEA Title I Evaluation Technical Assistance Program*. Submitted to the U. S. Department of Health, Education, and Welfare, Office of Education, April 1979.

Horst, D. P., Tallmadge, G. K., and Wood, C. T. *A Practical Guide to Measuring Student Achievement. Number 1 in a Series of Monographs on Evaluation in Education*, U. S. Department of Education, 1975.

Kaskowitz, D. H. and Norwood, C. R., *A Study of the Norm-Referenced Procedure for Evaluating Project Effectiveness as Applied in the Evaluation of Project Information Packages*. Prepared for the U.S. Office of Education, Office of Planning, Budge and Evaluation, January 1977.

Linn, Robert L. Measuring Pretest–Posttest Performance Changes. In R. A. Berk (Ed.) *Education evaluation methodology: the state of the art*, Baltimore, MD, The Johns Hopkins University Press, 1981.

Powers, Stephen, Slaughter, Helem, and Helmick, Chery. A Test of the Equipercentile Hypothesis of the TIERS Norm-Referenced Model. *Journal of Educational Measurement*, Volume 20, Number 3, pages 299-302, Fall 1983.

Public Law 100–297, the Augustus F. Hawkins–Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988, April 28, 1988.

Shepard, Laurie A. *Inflated Test Score Gians: Is It Old Norms or Teaching to the Test?* Final Deliverable, Prepared for the Department of Education by the University of California at Los Angeles Center for the Study of Evaluation, March 1989.

Sinclair, Beth, and Guttman, Babette. A Summary of State Chapter 1 Participation and Achievement Information for 1987–88. Prepared for the U. S. Department of Education, Office of Planning, Budget and Evaluation by Decision Resources, Washington, DC, August 1990.

Tallmadge, G. Kasten and Wood Christine T., User's Guide: ESEA Title I Evaluation and Reporting System. Prepared for the U. S. Department of Health, Education, and Welfare, Office of Education by the RMC Research Corporation, Mountain View, CA, October 1976.

Tallmadge, G. Kasten, An empirical assessment of the norm-referenced evaluation methodology. *Journal of Educational Measurement*, Volume 19, Number 2, pages 97-112, Summer 1982.

Tallmadge, G. Kasten. Rumors Regarding the Death of the Equipercentile Assumption May Have Been Greatly Exaggerated. *Journal of Educational Measurement*, Volume 22, Number 1, Pages 33-39, Spring 1985.

U. S Department of Education, *Chaper 1 Programs in Local Educational AGencies; Final Regulations*, 34 CFR Part 75 et. al., *Federal Register*, May 19, 1989.

U. S. Department of Education, Office of Elementary and Secondary Education, *Policy Manual*, April 1990.

2v