DOCUMENT RESUME

ED 244 343 EA 016 771

AUTHOR Keesling, J. Ward

TITLE Differences between Fall-to-Spring and Annual Gains in Evaluation of Chapter 1 Programs.

INSTITUTION Advanced Technology, Inc., McLean, Va.; Education Analysis Center for State and Local Grants (ED), Washington, DC.

SPONS AGENCY Department of Education, Washington, DC. Office of Planning, Budget, and Evaluation.

PUB DATE Jan 84

CONTRACT 300-82-0380

NOTE 35p.; Prepared for the Planning and Evaluation Service.

PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Educationally Disadvantaged; Elementary Secondary Education; *Pretests Posttests; *Standardized Tests; Testing Problems; Testing Programs; Test Norms; *Test Reliability; Test Results; Test Use; *Test Validity

IDENTIFIERS Education Consolidation Improvement Act Chapter 1; Elementary Secondary Education Act Title I; *Title I Evaluation and Reporting System

ABSTRACT

The Title I Evaluation and Reporting System (TIERS) was developed in order to examine the extent to which Title I (now Chapter 1) is remediating the disadvantages in basic skills of educationally deprived children. TIERS Model A contrasts the achievement of Chapter 1 students to publishers' norms for hypothetically comparable groups of students. However, the gains reported by districts using fall-to-spring testing cycles far exceed those of districts on the annual cycle (fall-to-fall or spring-to-spring). Three sources of problems in using Model A may account for this difference: (1) the norm tables of published tests may not be relevant to Chapter 1 students; (2) the publishers' norms may be used inappropriately; (3) local testing practices may bias the outcomes. Findings supporting each hypothesis are discussed indepth, leading to the conclusion that districts should adopt an annual testing paradigm, since fall-to-spring NCE (Normal Curve Equivalent) gains are unlikely to be accurate reflections of the true impact of Chapter 1. (TE)

DIFFERENCES BETWEEN FALL-TO-SPRING AND ANNUAL GAINS IN EVALUATION OF CHAPTER 1 PROGRAMS

J. WARD KEESLING

Advanced Technology, Inc.
7923 Jones Branch Drive
McLean, Virginia 22102

Prepared For:

Planning and Evaluation Service
U.S. Department of Education

JANUARY 1984

EDUCATION ANALYSIS CENTER FOR STATE AND LOCAL GRANTS

TABLE OF CONTENTS

LIST OF EXHIBITS

OVERVIEW AND SUMMARY

The Title I Evaluation and Reporting System (TIERS) was developed in order to examine the extent to which Title I (now Chapter 1) is remediating the disadvantages in basic skills achievement of educationally deprived children. Data collected via TIERS are intended to answer the question, "How much more did pupils learn by participating in the Title I project than they would have learned without it?" (Tallmadge and Wood, 1976a, p.2).

Most LEAs use TIERS Model A, which contrasts the achievement of Chapter 1 students to publishers' norms for hypothetically comparable groups of students. One clear piece of evidence that it may not always be appropriate to use publisher norms as the comparison is the large discrepancy between gains reported by districts using fall-to-spring testing cycles compared to those using annual testing cycles (fall-to-fall, or spring-to-spring) in Model A. At the national level the aggregated differences between the testing cycles are larger than the aggregated gains reported under the annual cycle. This is evidence of a strong method effect. This effect seems to be largely due to the fact that fall test scores are very low; the spring test scores do not seem to vary according to the testing cycle employed. This suggests that the true effects of Chapter 1 are similar no matter what the testing cycle and that the problem is confined to the fall testing.

Three sources of problems in using Model A are explored for the possibility that they account for exaggeratedly low fall scores (or higher spring scores) that would account for the difference observed between the two testing cycles. These problem sources are:

- The norm tables of published tests may not be relevant to Chapter 1 students.
- The publisher's norms may be used inappropriately.
- Local testing practices may bias the outcomes.

There are two ways in which the norm tables of published tests may not be relevant to Chapter 1 students: The samples of students used by publishers may not be

Note: This report was prepared pursuant to contract Number 300-82-0380, U.S. Department of Education. The technical monitor for this report was Dr. Robert Stonehill, U.S. Department of Education. The opinions and conclusions expressed in this report are those of the authors and do not necessarily represent the position or policies of the U.S. Department of Education.

representative of these students, and the curricula implicit in the tests may not correspond to the content of texts used to instruct Chapter 1 students. Baglin (1981) reports that very small fractions of the test publishers' initial samples of districts agree to participate in norming studies, and that acceptances are harder to obtain from large urban districts. Jaeger (1979) reported that different tests would report different Normal Curve Equivalent (NCE) gains for the same percentile change on a common scale. These effects were more pronounced for percentiles in the ranges served by Chapter 1.

Freeman, Kuhs, Porter, Floden, Schmidt and Schwille (1983) show that tests and texts show considerable variation in overlap. Linn et al. (1982) conclude, "Careful test selection and/or adjustments in the instructional materials to improve the match provides a project with a net advantage in comparison to the norms against which the gains are judged . . . ." A test that covers the curriculum to be taught would presumably show low scores in the fall relative to the norm (the students in the project will have less exposure to the content than the norm group because that content is yet to be taught in the project), and higher scores in the spring (because the project students will have more targeted instruction than the norm group).

Although it is hard to quantify these effects, the investigation reported here suggests that both the nonrepresentativeness of the samples and the variations in curricular overlap are likely to make published norm tables inappropriate to the assessment of Chapter 1 students when this assessment is conducted in a fall-to-spring testing cycle.

There are two primary ways in which test publisher norms can be used inappropriately. Errors can be made in converting raw scores to NCE scores, and interpolations to account for discrepancies between the actual testing date and the norming date can introduce biases.

Errors in converting raw scores to NCE equivalents may occur with some frequency, especially where the procedure is not automated. Linn et al. concluded that conversion errors could result in spurious gains of about 1 NCE in magnitude. Because fall-to-spring testing may involve more tables, conversion errors may combine to produce even larger spurious gains than in annual testing cycles.

Interpolation between the test publisher norming date and the actual date of testing is usually performed using an assumption that growth is linear between the fall empirical norm and the spring empirical norm. Evidence from a variety of sources reported in the paper shows that this assumption is not likely to be accurate. The result is that fall tests given before the norming date generally result in spuriously low NCE values. Testing early in the spring does not make up for the spurious deficit in the fall because the growth curve is steeper in the fall than in the spring. It is likely that this effect contributes about 2 NCEs to the difference between the two testing cycles.

Local testing practices can also have strong influences on the outcomes of Chapter 1 assessments. Several authorities mentioned "stakeholder effects" that would tend to make fall scores lower than anticipated and spring scores higher. On an annual testing cycle these effects probably balance out, but most would tend to exaggerate gains in a fall-to-spring testing cycle. Among these effects are:

- Not encouraging the best performance on fall pretests
- Increasing motivation to do well on the spring posttests
- Teaching test-taking skills
- Teaching specific test items
- Coaching during the posttest
- Selecting out low-scoring students at the postest
- Holding lower performing students back a grade

No published studies of these phenomena could be located that would permit estimation of the magnitude of these effects. One very carefully documented report from a larger school district did reveal that teaching test items can produce very large gains (21 NCEs in this particular case). It did seem clear, however, that most of these effects would be likely to contribute to the observed discrepancy between the gains reported using annual and fall-to-spring test cycles.

The more modest gains reported by projects using annual testing cycles correspond to the gains reported in other studies of the effects of compensatory education. Because these gains represent increments over-and-above the expected growth in basic skills achievement, even modest gains, if cumulated, can become important. For example, a student at the 25th percentile would be moved to the 35th by three years of exposure to a project that produced annual gains of 2 NCEs. It is important to continue to collect information via TIERS to document that Chapter 1 is capable of having this sort of impact.

The best advice to be given now is to repeat the conclusion of Linn et al. (1982) that districts should save money and testing burden by adopting an annual testing paradigm. Fall-to-spring NCE gains are unlikely to be accurate reflections of the true impact of Chapter 1.

DIFFERENCES BETWEEN FALL-TO-SPRING AND ANNUAL GAINS IN EVALUATION OF CHAPTER 1 PROGRAMS

BACKGROUND

The Title I Evaluation and Reporting System (TIERS) was developed in order to examine the extent to which Title I (now Chapter 1) is "working" to remediate the disadvantages in basic skills achievement of educationally deprived children. The system utilizes evaluation models developed by RMC (Tallmadge and Wood, 1978) under a contract from the United States Office of Education (USOE). This contract was a part of USOE's efforts to implement those sections of the Education Amendments of 1974 that required the Commissioner of Education to provide assistance to state departments of education to assist local educational agencies to develop and apply systematic methods of evaluation.

Data collected via TIERS are intended to answer the question, "How much more did pupils learn by participating in the Title I project than they would have learned without it?" (Tallmadge and Wood, 1976a, p.2). This question can be given a more formalized expression, utilizing a variation of a general model proposed by Rubin (1972), as follows:

> The effect on a particular student's achievement of participating in Chapter 1 supplementary programs versus participating only in the usual curriculum is the difference between: (1) the achievement score of the student at posttest if the student received Chapter 1 services (for a period of time), and (2) the achievement score of the student at posttest if the student received only the usual curriculum (during the same period of time).

Because an individual student must be assigned to either Chapter 1 or to the usual curriculum for a given period of time, this ideal model cannot be implemented. An alternative model, the randomized experiment, has been developed to provide a framework in which "the expected value of the difference in mean (achievement) scores . . . is equal to the average difference that would be observed if all (students could be exposed to both Chapter 1 supplements and to the usual curriculum alone) during the same time interval" (Linn and Slinde, 1977, parenthesized material added). TIERS Model B utilizes random assignment, but is very infrequently employed in actual evaluations of Chapter 1.

Most local education agencies (LEAs) use TIERS Model A, which contrasts the achievement of Chapter 1-served students to a hypothetically comparable usual-curriculum-only group. Comparability rests on the assumption that a Chapter 1 student would, if exposed to the usual curriculum only, remain at the same percentile rank among all students throughout their educational experiences. The national norms supplied by publishers of standardized tests are used to estimate the expectation under the usual-curriculum-only condition.

One clear piece of evidence that it may not always be appropriate to use publisher norms to estimate the usual-curriculum-only condition is the large discrepancy between gains reported by districts utilizing fall-to-spring testing cycles compared to those using annual testing cycles (either fall-to-fall, or more often, spring-to-spring) in Model A. Exhibits 1 and 2 are taken from Anderson (undated) and show the magnitude of these discrepancies.

These differences in gains systematically favor the fall-to-spring testing cycle and seem to be largely due to the differences in pretest scores: The fall tests yield lower scores than the spring tests. Since the posttests are relatively close in magnitude (except for the upper grades, which will have large standard errors as a function of the small numbers of projects operating at those levels), it seems that the spring results are quite similar and do not depend upon the testing cycle. Thus, the major issue is why the fall test scores are so low.

The differences in gains are far from trivial. The median difference is 3.8 Normal Curve Equivalents (NCEs) for reading, which exceeds any of the aggregate estimated reading gains from the annual cycle reports. The median difference of 5.7 NCEs for mathematics also exceeds the largest aggregate mathematics gain estimated from annual cycle reports. There is a powerful method effect at work in these data.

The remainder of this paper explores reasons why these two testing cycles produce such discrepant results. These reasons will be developed in the context of the Model A variation on the general model proposed above, and will reflect on the suitability of the assumptions utilized by Model A.

EXHIBIT 1. Differences Between Fall-to-Spring (FS) and Annual (AN) 1979-80 Title I Evaluation Results for Reading

	Weighted Normal Curve Equivalents									Weighted	
Grade	Pretest			Posttest			Gain			Number Tested	
	FS	AN	Diff.	FS	AN	Diff.	FS	AN	Diff.	FS	AN
2	30.8	37.6	-6.8	40.2	38.6	1.6	9.4	1.0	8.4	310,555	85,019
3	28.7	34.3	-5.6	36.1	36.7	-0.6	7.4	2.4	5.0	293,909	108,708
4	28.7	34.7	-6.0	35.6	36.6	-1.0	6.9	1.9	5.0	270,826	108,576
5	29.4	33.9	-4.5	35.5	36.2	-0.7	6.1	2.3	3.8	246,159	112,387
6	29.7	33.9	-4.2	35.7	37.2	-1.5	6.0	3.3	2.7	212,819	107,706
7	28.8	33.9	-5.1	34.3	35.8	-1.5	5.5	1.9	3.6	152,417	66,923
8	29.0	33.6	-4.6	34.0	35.8	-1.8	5.0	2.2	2.8	122,013	58,026
9	28.3	32.0	-3.7	33.5	33.8	-0.3	5.2	1.8	3.4	66,475	30,082
10	28.6	30.2	-1.6	32.8	29.5	3.3	4.2	-0.7	4.9	36,102	14,215
11	27.3	27.5	-0.2	30.5	25.3	5.2	3.2	-2.2	5.4	17,734	8,579
12	25.6	25.4	0.2	30.0	26.8	3.2	4.4	1.4	3.0	8,383	7,146

EXHIBIT 2. Differences Between Fall-to-Spring (FS) and Annual (AN) 1979-80 Title I Evaluation Results for Mathematics

	Weighted Normal Curve Equivalents									Weighted	
Grade	Pretest			Posttest			Gain			Number Tested	
	FS	AN	Diff.	FS	AN	Diff.	FS	AN	Diff.	FS	AN
2	32.0	41.9	-9.9	42.5	43.0	-0.5	10.5	1.1	9.4	124,576	50,084
3	31.5	39.7	-8.2	40.1	40.1	0.0	8.6	0.4	8.2	137,608	65,407
4	30.8	37.5	-6.7	39.8	39.2	0.6	9.0	1.8	7.2	147,338	70,637
5	30.5	36.6	-4.9	38.7	39.0	-0.3	8.2	2.5	5.7	136,872	71,038
6	30.9	35.4	-4.5	38.6	39.3	-0.7	7.7	3.9	3.8	119,003	69,002
7	30.6	34.5	-3.9	36.9	36.7	0.2	6.3	2.2	4.1	74,807	36,268
8	30.1	34.3	-4.2	36.3	37.1	-0.8	6.2	2.8	3.4	60,747	29,530
9	29.8	34.6	-4.8	35.9	35.1	0.8	6.2	0.5	5.7	28,579	15,971
10	32.0	32.9	-0.9	37.3	31.6	5.7	5.3	-1.4	6.7	12,192	7,718
11	32.5	34.9	-2.4	38.1	35.3	2.8	5.6	0.4	5.2	5,270	4,158
12	30.7	33.8	-3.1	37.2	34.9	2.3	6.5	1.0	5.5	2,195	3,587

In the Model A variation on the general model for evaluating Chapter 1, the growth of Chapter 1 students is contrasted to the growth of students in the publishers' norming studies who achieve at the same level at the time of the pretest. Three flaws can occur to make this an inappropriate comparison:

- the norm tables of published tests may not be relevant to Chapter 1 students
- the publishers' norms may be used inappropriately, and
- local testing practices may bias the outcomes.

Each of these will be discussed in turn and related to the phenomenon of fall-to-spring gains being larger than those from annual testing cycles.

RELEVANCE OF PUBLISHER NORMS TO CHAPTER 1 STUDENTS

Published norm tables for different tests may not have equal relevance to Chapter 1-served students. There are two reasons for this:

- the norming groups may not be representative of Chapter 1 students, and
- the tested curriculum may not be the curriculum that is taught.

The evidence to be presented indicates that publishers may not attain fully representative norming samples, and that there are considerable discrepancies in the implicit curricular content of standardized tests. Large differences in NCE gains can result from such variations in samples of students and content. Some of these variations may directly influence the difference between fall-to-spring or annual gain scores, while others may contribute to interactive effects that are discussed in a subsequent section.

Test publishers select districts for participation in their norming studies using probability sampling methods that would permit the construction of accurate national norms if all the selected districts agreed to participate. Baglin (1981) reports, however, that only 13 to 32 percent of the initially selected districts agreed to participate in recent norming studies, and that some publishers were unable to fill some of the sampling cells specified by their design. Baglin (personal communication, 1983) also

states that the publishers did have difficulty in persuading large urban districts to participate in norming studies (particularly those with enrollments in excess of 100,000 students). It is not clear that weighting the results can make up for missing one or more of these large districts. Under many reasonable sampling schemes the nation's largest districts would come into the sample with certainty and no amount of weighting could compensate for a refusal to participate.

Strand (Test Information Center, personal communication, 1983) indicates that her attempts to determine from test publishers what proportion of students participating in norming were served by Title I was unsuccessful. Thus it is hard to say whether the test publishers have represented the Chapter 1 population adequately in the norms. This could have serious consequences for Chapter 1 evaluations.

Suppose, for example, that norm-group students who achieve at the same level as Chapter 1 students on the pretest are not as likely to be economically disadvantaged and that they have higher rates of academic growth because of that difference. Over time, Chapter 1 students might not maintain the same percentile ranking because of the difference in growth rates. By itself, this effect might not have consequences for the difference between annual gains and fall-to-spring gains, but it may interact with other phenomena to produce some of those differences, as discussed in a later section of this paper.

If all nationally normed tests were equally appropriate to the Title I/Chapter 1 population of students, then one would expect that similar percentile gains (e.g., from the 10th to the 15th percentile) on all tests would register similar NCE gains. A major study of nationally-normed tests, the Anchor Study[1] (Loret, Seder, Bianchini and Vale,

[1] The Anchor Study used editions of tests that are now out of date. These tests were normed in an era when the acceptance of invitations to participate in norming studies was considerably higher than it is at present. For example, CTB/McGraw-Hill reported to the Technical Advisory Committee of the Systemwide Testing Program of the Department of Defense Dependents Schools (October, 1982) that 85 to 90 percent of their first choice districts participated in the 1968 norming of the CTBS Form Q, while only 15 percent of the first choices participated in the norming of CTBS Form V a decade later.

1974), compared the scalings of eight standardized reading comprehension tests, and concluded that the scales seemed generally comparable. However, Jaeger (1979) performed extensive secondary analyses of these data and concluded that identical percentile gains on the common scale derived in the Anchor Study would result in quite different NCE gains being reported for the eight different tests, especially for scores in percentiles below the 20th. Linn, Dunbar, Harnisch and Hastings (1982) were uncertain as to the meaning of the lack of national representativeness, although they cited work by Roberts that indicated that quite different NCE gains could result from different normative samples.

While it is not certain that the norm groups used by various test publishers vary in the extent to which they are representative of Chapter 1 students, the evidence is that such variation may exist, and is certainly important. It does seem likely that certain types of Chapter 1 students (those in large urban districts) may be underrepresented in norming groups, and this degree of underrepresentation may be increasing over time. The resulting bias in the estimated national growth rates for lower-achieving students could contribute to spurious assessments of the impact of Chapter 1.

Walker and Schaffarzick (1974) demonstrated that "students using different curricula in the same subject generally exhibited different patterns of test performance, and that these patterns generally reflected differences in the content inclusion and emphasis in the curricula." Wiley and Bock (1967) give a short example showing that very large differences in outcomes result from conscious choices to include or not include certain material in the curriculum. Tallmadge (1977) reviewed many other studies that showed that the content coverage of nationally normed tests varied widely. Wiley (1979a) asserted that variations in curricular content coverage could very easily mask other variations (e.g. pupil-teacher ratios, presence or absence of Chapter 1 funding) in instructional settings that might be the objects of evaluation. Leinhardt and Seewald (1981) proposed measures of curricular/test overlap to use in conducting research and evaluation.

More recently, Freeman, Kuhs, Porter, Floden, Schmidt and Schwille (1983) have demonstrated that popular textbooks and popular tests do not cover the same curricular content in fourth-grade mathematics. Freeman, Belli, Porter, Floden, Schmidt and Schwille (1983) have pursued this further to demonstrate that the manner in which the teacher utilizes the textbooks can also influence the degree to which the curriculum overlaps the test. The implications of this literature are that the choice of text and teaching method may have an important influence on the degree to which students are exposed to the curriculum implicit in the standardized test used to evaluate the outcomes of instruction.

It will be useful to illustrate how much of a difference this match can make in the content coverage. Using tabulations in Freeman, Kuhs, et al. (1983), it can be determined that if one were to use the Houghton-Mifflin textbook in fourth grade, the Iowa Test of Basic Skills (ITBS, also published by Houghton-Mifflin) would cover 42.9 percent of the topics to which the text devotes 20 or more problems. The Metropolitan Achievement Test (MAT) and the Stanford Achievement Test cover less than 31 percent of these topics. The CTBS Form S, Level II would cover nearly 47 percent of these topics, but this test only has spring norms, and according to the California State Department of Education (Test Planning Guide, 1982) the range of reliable measurements for this test extends from the 12th to the 92nd percentile, which does not cover the achievement range of many Chapter 1 students.

The district that chooses a test on grounds of tradition, cost, or because of a mandate from some other agency (e.g., the state) will find that judicious choice of text can make a large difference in the coverage overlap. Freeman, Kuhs, et al. (1983) show that the text published by Holt provides at least 20 problems each on 50 percent of the topics on the MAT, but only 22.2 percent of the topics on the Stanford. Interestingly, this is the maximum coverage provided for either test. As one might suspect, a district using the ITBS would be well advised to use the Houghton-Mifflin text as it provides at least 20 problems each on topics addressed by 31.8 percent of the tested items--the most of any text.

The User's Guide emphasizes the importance of selecting a test that matches the curriculum being evaluated, and with the availability of the literature cited above in additon to this encouragement, it is likely that test choices have tended to enhance the overlap between the two. Linn et al. (1982) conclude, "Careful test selection and/or adjustments in the instructional materials to improve the match provides a project with a net advantage in comparison to the norms against which the gains are judged. . . ."

It is hard to reach a firm conclusion, however, as to whether that advantage is more important to fall-to-spring testing cycles or to annual testing cycles. Baglin (1981) reports that test publishers found districts that were using their texts to be more willing to participate in norming studies. This means that the norms are perhaps slightly biased to reflect greater test-curriculum overlap than would be true in a strictly random sample, and the norm groups are possibly biased to the extent that some textbooks may be used by certain segments of the population more than others (which leads back to the question of representative norm groups discussed earlier). If these relationships were perfect, than we might be able to speak of "user norms" rather than national norms. One would not expect the test-curriculum overlap to create any problems if the national norms were truly "user norms" (except, perhaps, in aggregating results across projects).

Unfortunately, it is hard to find empirical evidence to demonstrate that higher than average test-curriculum overlap (relative to the national norm) will enhance NCE gains. One extreme example with very large gains is given later in the paper. Presumably if a test is given in the spring at the end of an instructional year in which the overlap has been higher than average, the posttest results should reflect a higher percentile standing. However, it is not clear that exposure to another year of higher than average overlap will produce the additional increment needed to make further NCE gains in a spring-to-spring annual testing cycle.

The same could hold true of fall-to-spring testing depending upon the level of the test. A fall test that covers the content of the previous year will reflect gains due to greater curricular overlap, and a subsequent year of instruction in content that overlaps the same test (used again in the spring) may not yield gains relative to the norm. However, choosing a test (to be used in fall and spring) that covers the to-be-taught curriculum might result in students scoring below the norm group in the fall and, with more exposure to the relevant curriculum during the year, scoring higher than the norm group in the spring.

Another possibility is that a test that is more sensitive to the curriculum might be sensitive to summer forgetting among students exposed to that curriculum. While most studies of summer gains or losses show that students tend to attain some growth in basic skills during the summer, the rate is quite a bit slower than the rate during the school year (Carter, 1980). Suppose that this slower rate of gain is a reflection of the loss of skills taught in school but unreinforced during the summer, and gains on other skills that are reinforced during the summer. A curriculum that is closely mapped to a particular test might result in students appearing to lose ground relative to the norm during the summer. The "saw-tooth" pattern of growth (Linn et al., 1982; Linn, 1981) may be exaggerated when the test used to measure growth is highly related to the curriculum used to instruct students. This could lead to higher than average fall-to-spring gains, while annual testing might produce little gain.

Clearly, the degree of overlap between test and curriculum is an important influence on achievement gains, and may account for a substantial part of the difference between annual and fall-to-spring gain scores. In combination with the evidence that norming groups may underrepresent Chapter 1 students, it appears that national norms for standardized tests may have only limited relevance to the evaluation of Chapter 1 students. At higher grade levels where Chapter 1 students are typically behind by several grade levels and may not be exposed to a curriculum at all like the one implicit in the tests, the norms could be much less relevant than those for younger students. This could explain the declining trend in NCE gains (especially true of fall-to-spring gains) from the lower to higher grade levels.

INAPPROPRIATE USE OF PUBLISHER NORMS

Test publisher norms are usually presented as extensive tabulations of conversions from raw scores to various other scales: percentiles, grade equivalents, stanines, expanded scale scores, and NCEs, to name some common scales. These tabluations are usually presented for specific periods of the year, so that testing accomplished within specific periods can be referred to the norm tables. Two flaws in the use of these tables can cause spurious NCE gains (or losses):

- conversion errors in which a table look up is performed incorrectly, and certain that conversion errors favor the fall-to-spring testing cycle. TIERS assessments of gains involve the contrast of two score averages no matter what the testing cycle. However, the fall-to-spring cycle probably involves the use of two different sets of tables, while an annual cycle could use only one, and this might increase the numbers of conversion errors.

Another source of the difference between fall-to-spring and annual gains is the fact that test publishers do not have empirical norms for all common testing dates in the fall and the spring. Older tests often had empirical norms only in the spring. Fall norms were created by interpolating between spring norms. The User's Guide is quite clear that tests must be used at times close to the publisher's empirical norm dates. Perhaps because of this strong insistance, most publishers now have both a fall and a spring empirical norming. Strand (1983) names the six tests most commonly used for Chapter 1 evaluations, and has indicated (in a personal communication) that all of these have both fall and spring norms. It should be noted that not all LEAs may be using these tests. As recently as the 1979-1980 school year, the State of California reported (Test Planning Guide, 1982) that 27 percent of schools in compensatory education programs were using tests with interpolated fall norms.

It is worth showing an example to indicate how much the use of interpolated fall norms can distort fall-to-spring gains. The data presented in Exhibit 3 come from a secondary analysis of data collected by the State of California in its evaluation of the Early Childhood Education Program (Burstein, Keesling, Conklin and Doscher, 1977). The tests involved are forms of the CTBS (Q, R, and S) that do not have empirical fall norms. The California State Department of Education interpolated (linearly) between spring norms to derive a single fall norm (set at October 15th).

EXHIBIT 3. The Influence of Pretest Date on Gains When Fall Norms Are Interpolated From Spring Norms.

	GRADE 1		GRADE 2		GRADE 3	
Month	Reading	Math	Reading	Math	Reading	Math
September	15.6	19.4	13.1	16.6	13.3	15.0
October	10.5	15.6	10.3	13.9	9.7	12.0
Difference	5.1	3.8	2.8	2.7	3.6	2.6

SOURCE: Burstein, Keesling, Conklin and Doscher, 1977. Table 3 (page 163) recomputed to show gains in NCE units. Nearly 100 schools tested in September and about 200 schools tested in October. At least 85 percent of these schools tested in April of the next year.

Testing at any time during September (which could be up to six weeks prior to the interpolated norming date) will result in a spuriously low pretest NCE score and a correspondingly inflated NCE gain score because September levels of achievement will generally be lower than October levels; indeed the linear interpolation model hypothesizes just this effect. As Exhibit 3 shows, the advantage of early testing amounted to at least 2.6 NCEs. Burstein et al. showed that interpolating exactly to the date of testing reduced these spurious gains, and accounting for slower growth rates over the summer (non-linear interpolation from spring-to-spring), reduced the differential gains even further.

As indicated earlier, the tests used most widely in Chapter 1 evaluations have empirical fall and spring norms. These norms mean that interpolations or projections can be made much closer to an actual data point, which should reduce the size of artificial gains. However, such artifacts are not entirely eliminated, as demonstrated below.

The User's Guide recommends that testing not occur more than two weeks before or after the publisher's norm date, but is not willing to declare test scores entirely out of bounds unless they are obtained six or more weeks away from the norm date. There is a variety of ways of dealing with the test data that arise from dates discrepant from the publisher's norm.

One state evaluation office (personal communication, 1983) indicated that they were using a canned computer package to process TIERS information that "threw out" any LEA report that involved testing more than a total of 30 days away from the published norms (adding together early testing in the fall and late testing in the spring). This system does, however, allow one to test 30 days early in the fall, and it compares all acceptable fall tests to the same norm, so that one can gain an advantage (spuriously low pretest score) from early fall testing.

A study by the California State Department of Education (Test Planning Guide, 1982) gives some indication of the problems that may be anticipated by testing too early or too late, and by using projected norms. Exhibit 4 condenses the results, which show that early fall testing and late spring testing can combine to yield a spurious gain of about 4 NCEs.

EXHIBIT 4. The Effects of Early and Late Testing, and the Use of Interpolated Norms for Reading Scores

Source of Effects	Effects (in NCE) on	
	Pretest	Postest
Using interpolated fall norms	-2	—
Early testing	-2	0
Late testing	+3	+2

SOURCE: Test Planning Guide published by the California State Department of Education. This table combines results for both fall-spring and annual testing cycles; the source does not report them separately. The source does not indicate the extent of early and late testing. The original tabulation was in percentile effects which have been converted to NCEs using the state average of 38 NCEs as the starting point.

Exhibit 5 presents more evidence of the effects of early fall testing. Even within the grace period recommended by TIERS it is possible to obtain an artificial loss in the fall of 3.7 NCE units. Scoring services provided by some publishers project norms for early fall testing (based on fall to spring growth) to the exact date of fall testing. They usually assume, however, that a linear growth rate is appropriate. Exhibit 4 shows that the difference between early and late testing is larger in the fall than it is in the spring, and it is not difficult to imagine that the actual growth curve of achievement might be like that shown in Exhibit 6. The linear projection or interpolation of norms based on fall and spring norming dates will lead to misrepresentations of the fall and spring NCEs and, consequently, the gains. These effects will probably inflate fall-to-spring gains, while they will not greatly influence annual gains if the annual testing occurs at the same time each year.

EXHIBIT 5. Computation of Spurious Losses Due to Early Fall Testing

CTBS Form S, Level B was normed in the cycle Spring-Fall-Spring. The reported means and standard deviations are:

Month	Raw Score Mean	Raw Score SD
April (0.7)	31.3	12.2
Nov. (1.2)	35.6	13.7
April (1.7)	59.4	18.4

Fall (Nov.) to spring (April) normal progress would be made at the rate of: 59.4-35.6/150 days = 0.16 raw score points per day. Using the fall standard deviation of 13.7, we can compute standard deviation units lost for each day testing occurs prior to the norm. For example: 30 days x 0.16 points per day = 4.8 points lost for testing one month early. This is equivalent to 4.8/13.7 = 0.35 standard deviation units or 0.35 x 21.06 = 7.4 NCE units.

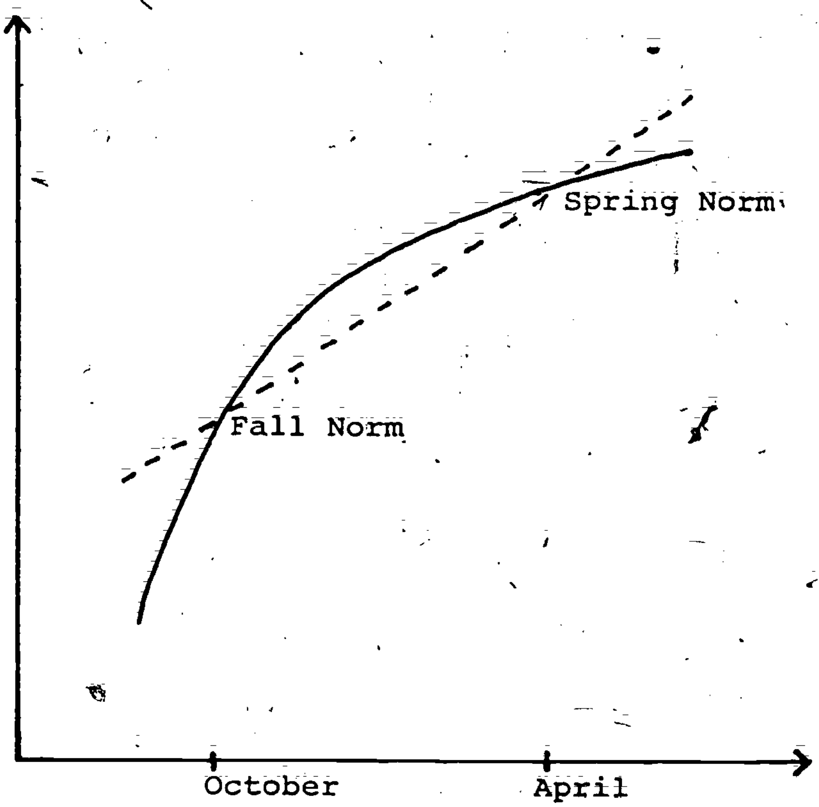
Computing these results for some likely testing dates yields:

Testing at	Which is	Produces a loss of
Mid September	6 weeks early	11.1 NCEs
Early October	4 weeks early[1]	7.4 NCEs
Mid October	2 weeks early[2]	3.7 NCEs

[1] Allowable under one reporting system if posttest is on norm date.
[2] TIERS recommended maximum gap in testing date.

SOURCE: Conklin, Burstein and Keesling, 1979.

EXHIBIT 6. Hypothetical Annual Growth Curve



The dashed line represents interpolations around the two norm dates (October and April).

A related question of importance is the incidence of early and late testing. The information in the Test Planning Guide (California State Department of Education, 1982) shows that in 1979-1980 evaluations, 16 percent of 2,527 schools evaluating compensatory education programs pretested early (by at least one day). Forty-eight percent (of 2,527 schools) tested late in the spring. Greater detail on testing dates was obtained from the Iowa State Department of Education (via personal communications, 1983). Data from this source are summarized in Exhibit 7.

EXHIBIT 7. Testing Dates for 1982-83 Evaluations of Chapter 1 in Iowa

POSTTESTING DATES

	Early					Late	
PRETEST DATES	at least 15 days	10 to 14 days	5 to 9 days	1 to 4 days	At Norm	1 to 4 days	5 or more days
15 or more days early	2	1	1	0	0	0	0
10 to 14 days early	1	19	4	1	0	1	0
5 to 9 days early	0	4	7	1	0	0	1
1 to 4 days early	2	1	7	9	1	6	1
At Norm	0	0	0	1	4	3	1
1 to 4 days late	0	1	1	6	0	2	1
5 or more days late	0	1	1	1	1	1	2
Total	5	27	21	19	5	13	6

The tabulation gives the percentage of 498 second grade reports (typically one per school) at each combination of pretest and posttest times. Rounding errors make the total add to 97 percent.

While the data in Exhibit 7 reveal that most schools are testing well within the TIERS recommended time limits, there is a clear bias in favor of early testing. Twenty-three percent test earlier in the fall than in the spring (the above-diagonal entries). They should show positive biases because the time elapsed between the testings is greater than the time between normings and because of non-linear growth as hypothesized in Exhibit 6. Thirty-seven percent test early by the same amount in fall and spring (the first four diagonal entries). If the model of Exhibit 5 is correct, then these cases will show a bias to spuriously low fall scores and, consequently, spuriously high gains, because the early pretesting is not fully compensated by early posttesting. Depending upon the shape of the curve some of the cases where the postest is earlier relative to the norm date than the pretest might still show the same bias because the pretest effect is much larger than the posttest effect. This means that there will be a bias to spuriously high gains in even these fall-to-spring testing cycles.

One of the explanations offered for early testing in the spring is that the schools want to be sure that they receive their results in time for the reports that are due to their respective state departments of education. Early fall testing is motivated by a desire to let teachers know more about the students they are teaching. There did not seem to be any reasonable explanation for the late testing in the fall.

Linn (1981) makes a strong case that more should be known about growth curves before it will be easy to compare the growth of one group of students against that of another. Having two norming points for most tests is simply not enough. Most of the studies that attempt to show that the norm group estimates of growth are reasonable proxies for the usual-curriculum-only treatment condition are based on fall and spring norming points within one year (Tallmadge, 1982; Powers, Slaughter and Helmick, 1983), or spring testing points over several years (Tallmadge and Fagan, 1977). To determine why fall-to-spring testing cycles appear to be biased we need more than two points on the growth curves for the students to be compared. Annual testing seems to have much better prospects for developing growth curves that will be useful in interpreting the nature of the gains made by Chapter 1 students. For example, Bock (1975) presents growth curves for vocabulary over four years that show different curves for high school males and females. Knowledge of such effects would be needed to properly assess the effects of special interventions such as Chapter 1.

We can now return to a point made much earlier in this paper. If the samples of students in the test publishers' national norm groups who score at the same fall pretest levels as Chapter 1 students have a different growth rate, then the curve for the norming sample that would be appropriate for Exhibit 6 may differ from the curve that would be appropriate for Chapter 1 served populations. This effect would be confounded with the problems of curricular overlap with the tests mentioned earlier: The growth curve of students exposed to curricula with greater overlap would be different from the growth curve of students exposed to other curricula. Furthermore, the growth curves for students at different initial percentile rank ranges might be different. This is important because some Chapter 1 projects are much more selective than others. Some states and districts only include students in the 25th percentile or below, while others include students below the 50th percentile. A further complication is the report of Mayer and Farnsworth (1983) that suggests that some students continue to grow at the previous rate after instruction has ceased, while others do not.

Clearly, a rather extensive study would be necessary to isolate all of these potential effects and prepare adequate growth curves. Ultimately, such a study might run into the difficulty that there would be so few students truly comparable to Chapter 1 students, but who are not receiving services, that it would not be possible to generate an expected growth curve under the usual-curriculum-only treatment condition. This would mean that the Chapter 1 effects would be included in the growth expectation for students at lower performance levels, and Model A would not be expected to detect gains relative to the norms.

If conversion errors contribute 1 NCE to the differential gains reported in Exhibits 1 and 2, and problems with linear interpolation contribute between 2 and 3 NCEs, the median difference is largely accounted for. It should be expected, however, that the effects of conversion errors or linear interpolation problems will interact with the problems of representative samples and content overlap. Conversion errors may occur at random, but their positive bias may mean that they occur more often when scores appear "too low." One incident was related in which all negative gain scores were converted to positive gains before the project report was submitted.

The shape of the growth curve (Exhibit 6) will surely depend upon the sample of students in the norm groups and the nature of the content match between the test and the curriculum. A single nationally-representative growth curve may only be an approximation to the actual situation in any local project.

LOCAL TESTING PRACTICES

In discussions with several authorities on testing in the preparation of this paper (TAC representatives, test publisher representatives, testing experts, state and local evaluators), a frequently expressed opinion was that testing in the fall and spring was performed under conditions different from those specified in the publisher's manuals. One of the interviewees called these differences "stakeholder effects." Stakeholder effects are different from the effects discussed earlier because they alter the degree to which Chapter 1 influences the testing results, while the others deal with the degree to which valid estimates of usual-curriculum-only treatment effects can be obtained. The effects discussed earlier will result in spurious gains whether or not there is a Chapter 1 project in operation; stakeholder effects generally augment any effect due to Chapter 1 with an effect of an additional treatment condition. When the effect of the additional treatment is not accounted for, Chapter 1 is credited with spuriously high scores.

Any alteration of the conditions for testing specified in the publisher's manual means that the publisher's norm tables are no longer valid. In general, the authorities contacted felt that deviations from the publisher's standardized conditions would produce lower fall scores and higher spring scores. The deviations from standard conditions that were mentioned include:

- Not encouraging the best performance on fall pretests (on annual cycles the pretest is also the posttest, and best performance is always encouraged)
- Emphasizing the importance of the posttest, increasing motivation to do well
- Teaching of test-taking skills
- Teaching specific test items
- Coaching during the posttest
- Selecting out low-scoring students at posttest
- Retention of lower performing students

Unfortunately, none of the authorities interviewed could provide a reference to any published (or fugitive) study of these phenomena. A check of all the listings for 1981 and 1982 in ERIC with the word "testing" in the title produced no likely entries. A check of the Current Index to Journals in Education from 1979 through 1983, under the headings "Testing Conditions" and "Testing Problems" also produced no relevant literature. The following compilation of tangential evidence and anecdotes gives a sense of the potential magnitude of the problem.

The basic premise of the stakeholder effect is that the fall testing will be done under conditions that tend to depress scores (or at least under conditions that do not raise scores beyond the effects of prior instruction), while the spring tests are conducted under conditions that will tend to raise scores. Annual testing cycles would result in no spurious gains if these effects occurred in each grade level. Fall-to-spring results would be strongly affected. For example, students probably know that the test in the fall is not important for their grade, or whether they will be promoted to the next grade level. Teachers probably tell them to relax and take it easy, that their scores will not matter. In the spring, however, the test is known to be important. It may determine promotion to the next grade. Teachers probably tell students that it is important and that they should try to do well. They probably encourage them to rest well the night before and eat well on the morning of the test. Would they do that for a fall testing?

One TAC representative suggested that fall testing is intended to identify students in need of services. Even though students may be taken to a separate area to be tested, and given a certain amount of encouragement, the purpose is to be sure to identify as many errors as possible in each test protocol so that 3 profiles of need can be developed. An LEA representative said, "The pretest was farcical. The objective was to qualify as many students as possible." In this LEA (which has since gone to annual testing), the time to complete various subsections in the fall was shortened from the publisher's recommendation, and the examiners would not clarify directions when asked..

Several of the people interviewed suggested that in the spring considerable attention is lavished on the preparation for the posttest. Instruction in Social Studies and Science begins to stress basic skills (sometimes these other subjects are not taught at all in the spring to make way for additional basic skills instruction). Teachers stress the parts of the curriculum they expect to be represented on the test, and they teach test-taking skills. It is arguable that explicit instruction in test-taking skills would constitute a special treatment, unlikely to be reflected in the publishers norms. It is also unlikely that teachers would devote much time to this subject in preparation for a fall test (although they should, if they want to obtain information about the subject content the students do not know, unconfounded with test-taking skills). Linn et al. (1982), citing work by Roberts, indicate that practice effects and instruction in test taking skills can have a sizeable influence on outcomes (several NCEs). Since these effects would tend to balance out on an annual testing basis, but are likely to be quite different in the fall than in the spring, they could be responsible for much of the difference in the NCE gains reported in the two testing cycles.

Probably the most obvious form of stakeholder effect is the deliberate teaching of items that will appear on the test. Achievement tests are usually composed of samples of items representing various skill domains. It is assumed that exposure to instruction will cover most of the domains to be tested and that the sampling of items will provide an accurate assessment of how much progress has been made on the entire set of domains. Emphasizing the instruction of specific test items in one or more domains will raise test scores, but probably means that the range of those domains has not been adequately covered.

It is important to distinguish this effect from that of choosing a test that emphasizes the type of problem found on the test. In the latter case one is emphasizing the overlap of the domains taught and tested, and while this can invalidate publisher's norms (for reasons discussed earlier), it is not the same as emphasizing instruction in the exact items to be tested. Maximizing the overlap between domains taught and tested should assure good coverage of the domains, while teaching to the specific test items limits the scope of coverage.

A good example of this phenomenon has been provided by Stephen Isaac of the San Diego Unified School District (personal communication). Under a court order to raise the achievement of students in minority-isolated schools, the district prepared a mastery learning project in basic skills. Evaluators in the district were conscious of potential problems with the security of the tests they planned to use in the evaluation (CTBS Form S) and eventually discovered that systematic instruction in 30 out of the 40 vocabulary items on the test had been offered to third grade students during the year. Each of these 30 items was included in a set of "Word Warm-Up Exercises" that were used to start the reading lessons. The stems and responses had been reversed (probably to hide this test-specific instruction). In addition to this use in instruction, 6 of 7 of these items were also used (with stems and responses still reversed from the CTBS format) in a series of "cumulative tests" given to all children. These tests were returned to the children so that they could learn the correct associations. It is estimated that the Word Warm-Ups provided direct teaching of the 30 items 3 times each during the year, including one exercise that was presented in a format identical to the CTBS test, except for the stem/response reversal.

Results reported for the CTBS Form S vocabulary testing showed a gain from the 33rd percentile (NCE=41) the previous spring (there was no prior item-specific instruction) to the 72nd percentile (NCE=62). The students were able to answer 12 more items correctly than before. Because this test-specific instruction was detected, Form T of the CTBS was given soon after Form S and students scored at the 43rd percentile (NCE=46) in vocabulary. Clearly, teaching to the test invalidates the publisher's norm tables as a means of determining the expected growth curve.

Teaching specific test items year after year would not yield large NCE gains on an annual basis, but would produce large NCE gains in a fall-to-spring testing cycle. Telling students the correct answers during the test session will have similar effects.

Another example of a stakeholder effect is in the selection of students to take the tests. Some LEAs may eliminate the scores of some low-achieving students from the posttest, and therefore, from the TIERS reports. California, for example, permits teachers to exempt limited-English proficient (LEP) students from testing in the English language. This may be a perfectly valid reason to protect some students from discouraging experiences, but it can be abused by withholding students who should be tested (i.e. are not truly LEP), but might make the project look ineffective. This form of student selection bias may not contribute to differentiating between the two testing cycles.

Another student selection device that could differentially affect outcomes under the two testing cycles would be to implement a strong policy of retention. Such a policy would be unlikely to be reflected in the publisher norms and would result in a number of effects that would boost test scores:

- The retained students would score low in the fall (perhaps even lower than the status that led to retention would lead one to predict, because they might be depressed at being retained).
- The retained students would score higher in the spring because they would have had another year's practice on the material (and they might be motivated to do well).
- The students sent on would be those who grew faster and might be likely to do so again. This would boost fall and spring scores.

Retention might benefit LEAs on fall-to-spring testing cycles.

CONCLUSIONS AND ADVICE FOR LOCAL DISTRICTS

Longitudinal studies tracking students for more than one year show that fall-to-spring gains are not maintained (see Linn et al. 1982; Linn, 1981; and Perry, 1983 for examples). TIERS reports of fall-to-spring gains appear to be too large. Linn et al. (1982) report that major studies of the impact of Title I have shown gains of about 1 or 2 NCEs per year in reading. The data from annual evaluation cycles reported in TIERS tend to support this degree of gain. Math gains appear to be a little larger than this.

Much of the evidence we have presented in this paper tends to indicate that the fall data point is more questionable than the spring data point. Interpolations or projections around the fall data point are more sensitive than are interpolations around the spring data point (because the growth curve is steeper in the fall). There is probably more variation in testing practices associated with fall tests than with spring tests. While the spring test is probably more generally played up as important regardless of whether one is in a fall-to-spring cycle or an annual cycle, the fall test may be treated quite variably.

In the metric of raw scores (number of items answered correctly), and possibly in expanded scale score metrics, the difference between fall and spring scores would reflect the actual amount of learning. Unfortunately these metrics are not comparable from test to test. NCE scores are intended to show the incremental gain over and above expectation that can be attributed to compensatory education. But, using NCEs to measure gains from fall-to-spring requires several assumptions about the nature of growth curves and the willingness of LEAs to use standardized testing practices that may not be realistic.

Because NCEs measure incremental effects, even small values are potentially important, especially if they cumulate. Carter (1980) shows evidence (p. 152) that about 60 percent of Title I students in a given year are in the program the next year also. Gains of 2 NCEs per year, cumulated for three years, would move a student from the 10th percentile at the end of Grade 1 to the 16th by the end of Grade 4. The same gain would move a student from the 15th percentile to the 23rd, and a student at the 25th percentile would be moved to the 35th. These are respectable gains.

Even though small LEAs (most LEAs in the country are of this size) will have too few served students to reliably detect such small gains, their data is needed in the TIERS system to document that Chapter 1 is producing effects of this magnitude on a nationwide basis. The advice to such small LEAs would be not to regard any one year's results as meaningful for local policy setting. An accumulation of data over time might prove more useful, although the large standard errors of the outcome measures may make it difficult to detect effects of changes in the nature of the program (such as the materials used, the types of teachers and aides employed) and these effects could be overwhelmed by any changes in the test used to assess outcomes.

Larger LEAs who switch from fall-to-spring testing to annual testing will probably wonder how to handle reporting lower gains. If they had been testing fall-to-spring for some time, they probably have noticed that they reported very large gains each year, but that the same students were in Title I (or Chapter 1) over the years and their cumulated gains were nothing like the sum of the yearly gains. Data that show the drop from spring to fall, but show that fall-to-fall there is some maintenance of gains (e.g., Perry, 1983), could probably be recovered from such testing systems to show that the annual testing cycle will give a better estimate of the gain that is likely to cumulate through time.

Apparently, some LEAs have asked for a way to estimate the fall-to-spring gain that would correspond to the annual gain they are estimating so that they can report better-sounding news to their school boards and parents. They should be advised to be more straightforward with these constituencies and use the method presented above to indicate why the new (annual) cycle will provide better information for policy purposes. In the interviews conducted with local evaluators in preparing this report, those in LEAs that had changed to the annual cycle indicated that the savings of money and burden on students and teachers far outweighed problems with reporting the data.

The best advice to be given now is to repeat the conclusion of Linn et al. (1982) that districts should save money and testing burden by adopting an annual testing paradigm. Fall-to-spring NCE gains are unlikely to be accurate reflections of the true impact of Chapter 1.

REFERENCES

Anderson, Judith. Differences between academic year and calendar year Title I evaluation results. Washington, D.C.: Undated draft.

Baglin, Roger F. Does "nationally" normed really mean nationally? Journal of Education Measurement, 1981, 18, 97-108.

Bock, R. Darrell. Multivariate Statistical Methods in Behavioral Research, New York: McGraw-Hill, 1975.

Burstein, Leigh, J. Ward Keesling, Jon Conklin and Mary Lynn Doscher. Auditing large-scale evaluation: The quality of evaluative information for the assessment of program impact and for decision-making. Studies in Educational Evaluation, 1977, 3, 155-168.

California State Department of Education. Test Planning Guide: Suggestions For Avoiding Testing Problems. Sacramento, CA: Author, 1982.

Carter, Launor F. The Sustaining Effects Study: An Interim Report. Santa Monica, CA: System Development Corporation, 1980. TM-5693/200/00.

Conklin, Jonathan E., Leigh Burstein and J. Ward Keesling. The effects of date of testing and method of interpolation on the use of standardized test scores in the evaluation of large scale educational programs. Journal of Educational Measurement, 1979, 16, 239-246.

Freeman, Donald J., Gabriella M. Belli, Andrew C. Porter, Robert E. Floden, William H. Schmidt, and John R. Schwille. The influence of different styles of textbook use of instructional validity of standardized tests. Journal of Educational Measurement, 1983, 20, 259-270.

Freeman, Donald J., Therese M. Kuhs, Andrew C. Porter, Robert E. Floden, William H. Schmidt and John R. Schwille. Do textbooks and tests define a national curriculum in elementary school mathematics? The Elementary School Journal, 1983, 20, 501-513.

Horst, Donald P. Checklists of potential errors in the ESEA Title I Evaluation and Reporting System. In: Bessey, Barbara L. (Ed.), Further Documentation of State ESEA Assistance State ESEA Title I Reporting Models and their Technical Assistance Requirements, Phase II, Volume II. RMC Report UR-331. Mountain View, CA: RMC Research Corporation, 1978.

Jaeger, Richard M. The effect of test selection on Title I project impact. Educational Evaluation and Policy Analysis, 1979, 1(2), 33-40.

Leinhardt, Gaea and Andrea Mar Seewald. Overlap: What's tested, what's taught? Journal of Educational Measurement, 1981, 18, 85-96.

Linn, Robert L. Validity of inferences based on the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, 1979, 1(2), 23-82.

Linn, Robert L. Discussion: Regression toward the mean and the interval between test administrations. New Directions for Testing and Measurement, Number 8: Measurement Aspects of Title I Evaluations, San Francisco: Jossey-Bass Publishers, Inc., 1980.

Linn, Robert L. Measuring pretest-posttest performance changes. In: Berk, Ronald A., (Ed.), Educational Evaluation Methodology: The State of the Art. Baltimore: Johns Hopkins University Press, 1981.

Linn, Robert L., Stephen B. Dunbar, Delwyn L. Harnisch and C. Nicholas Hastings. The validity of the Title I Evaluation and Reporting System. Chapter Two in: Assessment of the Title I Evaluation and Reporting System. Elizabeth R. Reisner, Marvin C. Alkin, Robert F. Boruch, Robert L. Linn, Jason Millman. Washington, D.C.: U.S. Department of Education, April 1982.

Linn, Robert L. and Jeffrey A. Slinde. The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 1977, 47, 121-150.

Loret, P.G., A. Seder, J.C. Bianchini and C.A. Vale. Anchor Test Study: Equivalence and norm tables for selected reading achievement tests (grades 4, 5 and 6). Washington, D.C.: U.S. Government Printing Office, 1974.

Mayer, Victor J. and Carolyn H. Farnsworth. The presence of a momentum effect in intensive time-series data on learning. Paper presented at the annual meeting of the American Educational Research Association, Montreal, 1983.

Perry, Marcia D. A meta-analysis of Title I/Chapter 1 sustained effects study. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, 1983.

Powers, Stephen, Helen Slaughter and Cheryl Helmick. A test of the equipercentile hypothesis of the TIERS Norm-Referenced Model. Journal of Educational Measurement, 1983, 20, 299-302.

Roberts, A. O. H. Regression toward the mean and the regression-effect bias. New Directions for Testing and Measurement, Number 8: Measurement Aspects of Title I Evaluation. San Francisco: Jossey-Bass Publishers, Inc. 1980.

Rubin, Donald B. Estimating causal effects of treatments in experimental and observational studies (ETS RB 72-39). Princeton, N.J.: Educational Testing Service, 1972.

Strand, T. Memorandum to Sustained Achievement Study Committee, Evanston, Illinois: Educational Testing Service (Test Information Center), October 12, 1983.

Tallmadge, G. Kasten. Title I evaluations: comparable outcome measures for dissimilar instructional treatments? Paper presented at the 27th Annual Conference of Directors of State Testing Programs, Princeton, New Jersey, October 1977.

Tallmadge, G. Kasten. An empirical assessment of norm-referenced evaluation methodology. Journal of Educational Measurement, 1982, 19, 97-112.

Tallmadge, G. Kasten and Barbara M. Fagan. Cognitive growth and growth expectations in reading and mathematics. RMC Report UR-326. Mountain View, CA: RMC Research Corp., November 1977.

Tallmadge, G. Kasten and Christine T. Wood. User's Guide: ESEA Title I Evaluation and Reporting System. Mountain View, CA: RMC Research Corp., October 1976a.

Tallmadge, G. Kasten and Christine T. Wood. Characteristics of Eight Commonly Used Nationally Normed Tests. Technical Paper No. 5, ESEA Title I Evaluation and Reporting System. Washington, D.C.: U.S. Office of Education, October 1976a.

Tallmadge, G. Kasten and Christine T. Wood. User's Guide: ESEA Title I Evaluation and Reporting System. Mountain View, CA: RMC Research Corp., Oct. 1978.

Walker, Decker F. and Jon Schaffarzick. Comparing Curricula. Review of Educational Research, 1974, 44, 83-112.

Wiley, David E. Policy-responsive evaluation. In: Quellmalz, Edys and Eva Baker (Eds.), Proceedings of the 1978 CSE Measurement and Methodology Conference. Los Angeles: University of California, 1979a.

Wiley, David E. Evaluation by aggregation: social and methodological biases. Educational Evaluation and Policy Analysis, 1979b, 1(2), 41-45.

Wiley, David E. and R. Darell Bock. Quasi-experimentation in educational settings: Comment. The School Review, 1967, 75, 353-366.

Wood, Christine T. Test norming practices and the norm-referenced evaluation model. In: Bessey, Barbara L. (Ed.), Further Documentation of State ESEA Title I Reporting Models and their Technical Assistance Requirements, Phase II, Volume II. RMC Report UR-331. Mountain View, CA: RMC Research Corp., 1978.

Wood, Christine T. and G. Kasten Tallmadge. Local Norms. Technical Paper No. 7, ESEA Title I Evaluation and Reporting System. Washington, D.C.: U.S. Office of Education, October 1976.