SOME METHODOLOGICAL ISSUES IN USING STANDARDIZED TEST SCORES TO EVALUATE LARGE-SCALE EDUCATIONAL PROGRAMS

Jon Conklin

CSE Report No. 111
March 1978

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

The research reported herein was supported in whole or in part by the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

SOME METHODOLOGICAL ISSUES IN USING STANDARDIZED TEST SCORES TO EVALUATE LARGE-SCALE EDUCATIONAL PROGRAMS[1]

Jon Conklin

In evaluating the impact of educational programs, the most commonly employed criterion is academic growth. Evaluators are often interested in ascertaining the degree and kind of achievement gains that are directly attributable to program effectiveness. To evaluate effectiveness in many non-experimental studies, standardized achievement tests are administered before and after program implementation, gain scores are computed, and comparisons are made to a norm group. Any deviations from "expected" gains are then attributed to program impact. This approach is widespread and often decisions on funding, program expansion, and program termination are based on results from this simple straight-forward procedure.

There are, however, serious inherent limitations in using such an approach that may lead to erroneous inferences and resulting policy decisions. For example, Beggs and Hieronymous (1968) have reported evidence that raises doubt as to the validity of the assumptions of uniform growth over the school year and that one month of growth occurs during the three summer months. David and Pelavin (1977) question the practice of comparing the growth of low achieving program participants to the growth expected in the norm population. They note that published norms are based on the achievement of the 50th

[1] Paper presented at the annual meeting of the California Educational Research Association, Marina Del Ray, California, November 17, 1977.

percentile student and that patterns of growth for that student are likely to be different from those for low achievers. In fact, a significant increase over the expected growth for low achievers may still fall short of the growth expected for the average achiever.

Ideally, the norm group should be a representative sample of the population from which the treatment group is drawn, and thus, disadvantaged children should be compared against a disadvantaged norm (Tallmadge & Horst, 1974). Other problems arise when academic gains of low achievers, which may partially be due to regression artifacts (extremes regress toward the mean), are wholly attributed to program impact. The use of gain scores to represent and compare growth has been criticized by Cronbach and Furby (1970) because such scores are usually negatively correlated with initial ability, partially due to this regression phenomenon.

Another potential problem concerns the temporal gap between pre- and posttesting. DeVito and Long (1977) point to several limitations in crossing test levels when evaluating growth due to longitudinal program impact, but still stress the importance of spring-spring testing as opposed to fall-spring testing because the latter generally inflates apparent gains. On the other hand, David and Pelavin (1977) argue that the only meaningful measure of sustained program impact in terms of academic growth is obtained through fall-fall testing.

We acknowledge the inherent limitations in comparing academic growth of program participants to expected growth in the norm population, but also recognize the widespread use of such comparisons. Therefore it seems appropriate to examine critically the actual procedures used to calculate and represent academic growth. The procedures considered specifically are those most typical of large-scale evaluation studies where neither the test nor the testing dates can be controlled. Few schools involved in a large-scale program actually test on days for which published norms are appropriate. In addition, for many of the tests used, published norms are based on only a single standardization during the spring of the school year. Thus fall norms must be estimated. The problems raised by these issues of varied testing dates and estimated norms in terms of the computation and use of standardized gains are of central interest here.

PROBLEM

For the sake of comparability, the raw scores from standardized tests given to different groups of students at different points in time are usually transformed to some common scale. A typical procedure is to transform all raw scores to standard scores by using norms based on publishers' standardization samples (e.g., the California State Department of Education evaluation of California schools receiving compensatory funding in 1974-1975 and 1975-1976: see California State Department of Education, 1976, 1977; see also Keesling & Burstein, 1977). Usually the raw score is deviated from the norm mean, divided by the norm standard deviation, and converted to a standard scale (in this study we use the T-scale with mean of 50 and standard deviation of 10). When available, fall scores are transformed using fall norms and spring scores are transformed using spring norms. Gains are then defined as the difference between the fall and spring standard scores.

In using this approach to represent growth, certain assumptions must be implicitly made. Most basically, the approach assumes that both the fall and spring standard scores are equally meaningful and valid; that is, that both use norms based on equivalent standardization samples. In reality, however, only spring norms have been established for most tests. Though this creates little problem for spring-spring testing,it is problematic in the case of the more common fall-spring testing schedules. Without explicitly measured fall norms against which to compare fall test scores, most investigators rely on linear estimation techniques. Test publishers often use these techniques to provide derived fall norms with limited explanation of the implications of this practice for fall-spring test comparisons.

To estimate fall norms linearly for, say, grade 3 the spring norms for grade 2 and grade 3 are plotted and fall norms are interpolated based on a ten-month year (one month growth is assumed for the three summer months). This procedure usually involves equating different test forms or levels, or assumes two cohorts different in grade levels to have the same growth pattern, and may be criticized on these bases alone (see DeVito & Long, 1977).

Unfortunately, linear estimation has another serious problem. Namely, it assumes that growth from spring to spring is linear, and that the fall norms (both means <u>and</u> standard deviations) can be estimated by taking the midpoint of this growth line.

Additionally, there is the problem of using norms to transform raw scores from testings that take place on dates other than those for which the norms are appropriate. Since academic growth is going on during the whole school year, a norm based on a March 15 standardization is not appropriate for use in transforming a May 15 score; i.e., it does not represent the true "expected" level of achievement for that date. However, a common practice is to use the norm means and standard deviations as fixed constants. Take for example, the California State Department of Education (SDE) evaluation mentioned previously. In SDE's transformation procedure, for a common test, schools pretesting in September, October, November, and December would have their scores adjusted by the same fall mean and standard deviation. Similarly, schools pretesting in March, April, May, or June would have their scores adjusted by the same spring mean and standard deviation. This is in spite of the fact that if continuing growth is assumed, the same values can not be the appropriate norms for differing test dates.

The consequences of the practices described above may be serious. First, evidence based on published norms of tests which have been explicitly standardized in both the fall and the spring indicates that linear estimation typically overestimates the rate of academic growth from spring of the previous year to fall of the current year even after subtracting two months for the summer.[2] Using these interpolated norms to transform fall test scores, then, results in spuriously low standard scores. Thus, gains from fall to spring are spuriously high. An example of this phenomenon is made apparent in Figure 1. This means that schools reporting fall-spring test data may look exceptionally good in comparison to the norm, whereas schools reporting spring-spring data are typically penalized, in a relative sense, because they fail to receive the spurious advantage associated with linear interpolation.

[2] In the case of the CTBS (Comprehensive Test of Basic Skills), first grade fall and spring norms are based on explicit standardization. Published values show a fall mean of 35.6 and standard deviation of 13.7; a spring mean of 59.4 and standard deviation of 18.4. The interpolated values for the fall mean and standard deviation are 45.4 and 15.3, respectively. (Previous spring values were 31.3 and 12.2)

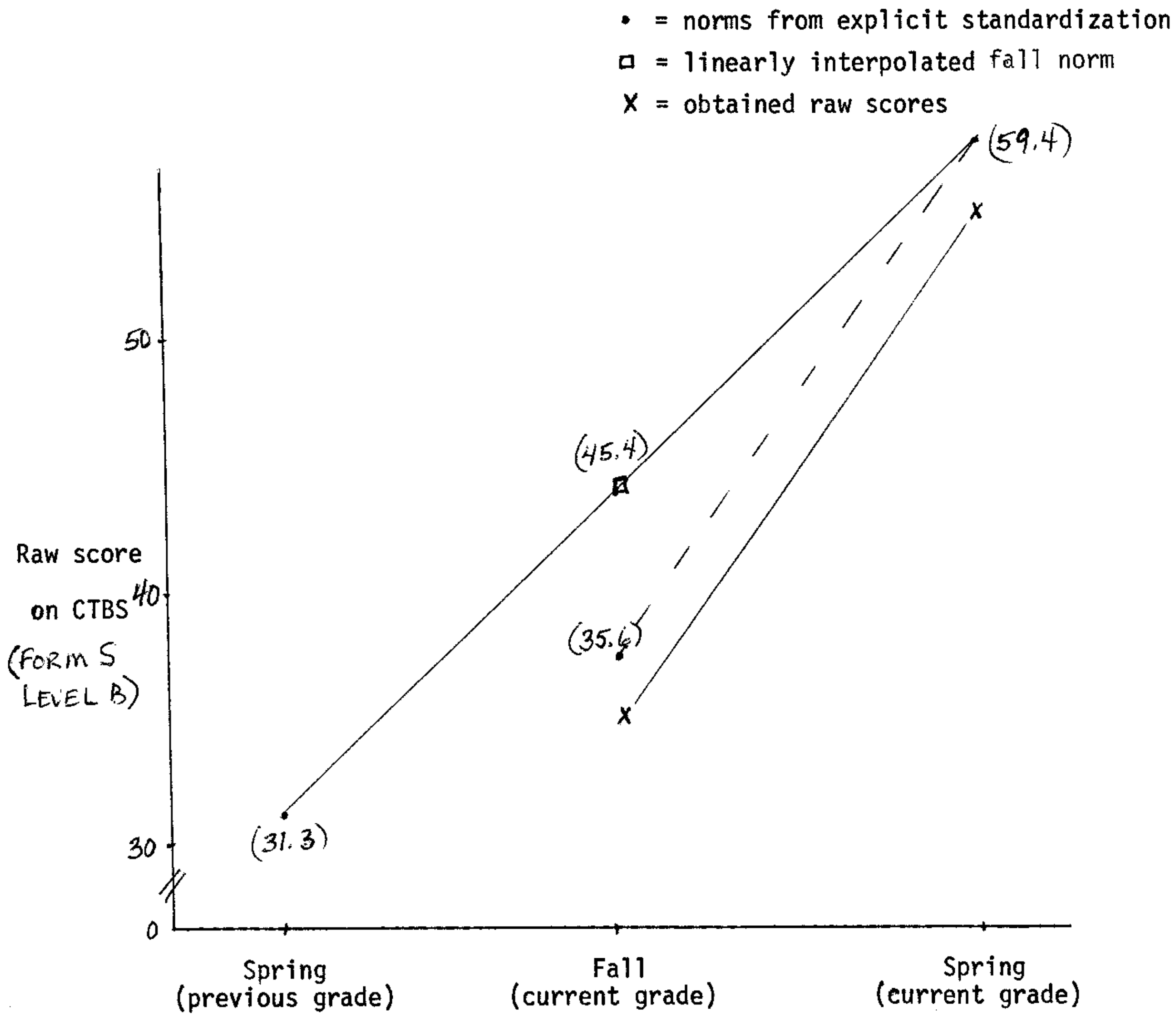


FIGURE 1 Published norm means for CTBS grade 1 (Form S Level B) are plotted along with the estimated fall norm mean based on linear interpolation and with a student's hypothetical raw scores from fall and spring testing. Whereas the student's position relative to the published norms stays constant, it changes when the estimated fall norm mean is used. Representing growth as the slope between the fall and spring values it can be seen that if we are looking at the published explicit norms his growth is identical to that of the norm group. On the other hand, if we look at the line between the estimated fall norm and the explicit spring norm his growth appears to greatly exceed that of the norm group.

Second, by using fixed dates for transforming fall and spring data, schools testing in early fall and late spring may appear to have greater gains than schools testing later in the fall or earlier in the spring. This is because comparisons made from many different test dates to one fixed norm fail to take into account any growth that takes place between those separate dates and the date for which the norm was established. Take, for example, the case in which the fall test norms are based on a mid-October standardization and a school has tested its students in mid-September. Though the actual expected norms for mid-September are lower than the norms for mid-October, the students' raw scores are transformed to standard score form using the published mid-October norm. The effect is that the students' pretest standard scores are spuriously low, and thus their fall-spring gains are spuriously high. A similar effect occurs with late spring testing where spuriously high posttest standard scores are obtained. A graphic display of this phenomenon is provided in Figure 2. It would appear that using linear estimation techniques to obtain date-appropriate derived norms would be preferred (albeit the problems of linear interpolation) over the use of norms as fixed values for transforming raw scores to standard scores and calculating gains.

Tallmadge and Horst (1974) have discussed these problems in A Procedural Guide For Validating Achievement Gains in Educational Projects. They emphatically state that interpolated norms, "while possibly useful for counseling or diagnostic purposes, are likely to be in error by amounts large enough to invalidate any inferences drawn about cognitive growth," and urge that such norms never be used to assess educational program impact. In addition, they point out that a treatment group should be tested at times exactly corresponding

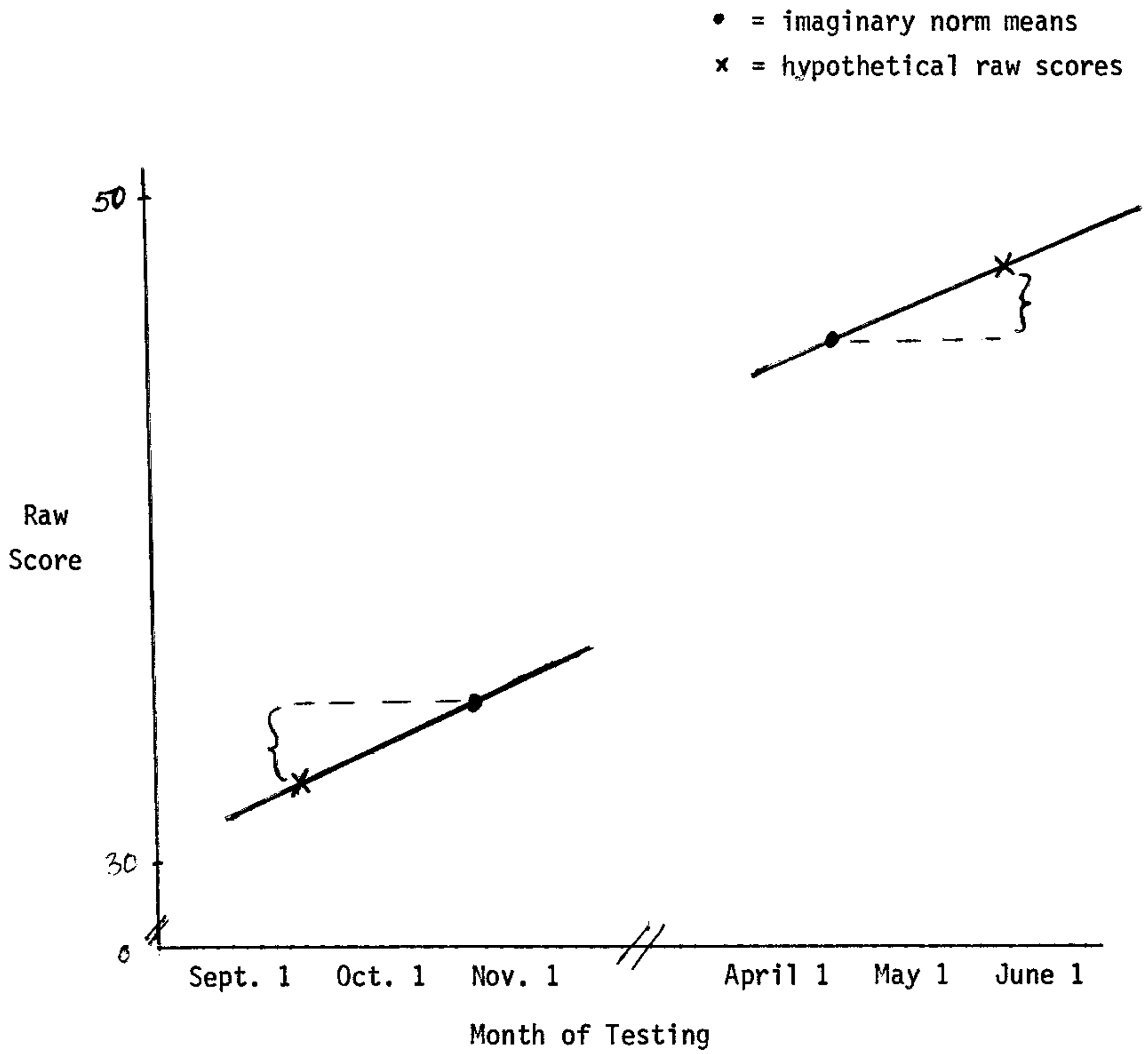


FIGURE 2 Imaginary fall and spring norm means are presented with lines depicting expected growth and with a student's hypothetical raw scores from fall and spring testing. Whereas the student achieves at the expected norm level for the dates on which he is tested, his scores relative to the fixed fall and spring norms are lower than "expected" in the fall and higher in the spring. Though the student has, in reality, grown exactly as the norm group, the apparent growth based on transformed scores using fixed norms is much greater than than of the norm group.

to real normative data points. Their arguments, though strong, still lack empirical examples. It is the purpose of this study to provide such examples.

To investigate these issues with real data and to examine their possible effects on evaluation conclusions or policy decisions, we apply several transformation procedures to existing test data. The focus here is limited to fall-spring testing in primary school grades and to students receiving the same form and level of a popular standardized achievement test for which only spring norms have been explicitly measured. Three transformation procedures are compared on the basis of their effects on pretest standard scores and the resulting pre- to posttest gains. The first is the commonly used one in which standard scores are based on linearly estimated and fixed norms. The second makes use of linearly estimated fall norms but adjusted fall and spring norms (by a single linear function) according to the date of testing. In the third procedure the fall norms are based on a hypothetical growth pattern evidenced by explicit fall and spring standardizations in an earlier grade and the procedure uses two linear functions (one for before and another for after the fall norm date) to adjust fall and spring norms according to the date of testing. If the arguments we have already made are valid, the use of adjusted norms will eliminate the spurious advantage of early pretesting and late posttesting, and the use of "curve-estimated" rather than linearly estimated fall norms will result in higher pretest standard scores and lower pre- to posttest gains.

METHOD

Data. The data used in this study were made available by the California State Department of Education in conjunction with an audit of their evaluation activities undertaken by the UCLA Center for the Study of Evaluation (Keesling & Burstein, 1977) and consist of second grade school means for pre- and posttesting during the 1974-1975 academic year. As part of an ongoing evaluation activity, the Department of Education requires California schools receiving compensatory funding to administer pre- and posttests using published standardized achievement tests. However, exact dates of testing and specific tests to be used are not mandated. Thus schools are free to test when they want and with the test they choose. In 1974-1975 the most widely used test was the CTBS (Comprehensive Test of Basic Skills) which was used by 25% of the reporting schools. In the second grade, standardization has been carried out only in the spring for the CTBS. Of the test results made available, the schools with students taking appropriate forms and levels of the CTBS reading and math tests during fall of 1974 and spring of 1975 totaled 385 for second grade reading and 361 for second grade math. The raw test means provided by these schools were used in our analysis.

Procedure. Linear estimation of fall norms proceeded as mentioned above. The published explicit spring norms (means and standard deviations) for grade 1 and grade 2 were plotted, a line was drawn between them, and the midpoint was located. The obtained values matched the grade 2 fall norms published in the CTBS manual. Since the actual standardization had been carried out on or about April 1, the derived fall norms were actually "appropriate" (assuming linear growth) for November 1. For two of the tranformation procedures examined, these linearly estimated fall norms were used. However, there is evidence that actual growth proceeds at a slower rate from spring of year 1 to fall of year 2 than the rate from fall of year 2 to spring of year 2 (Beggs & Hieronymous, 1968; David & Pelavin, 1977). In hopes of providing some bounds on the size of possible gains, we will base our estimation of second grade fall norms in the third procedure on the pattern of growth shown in the first grade (for which CTBS does explicitly standardize in fall and spring). For grade 1 the exact linear functions from grades 0.7 (previous spring) to 1.2 (current fall) and from grades 1.2 to 1.7 (current spring) could be determined. For the third procedure, the separate functions for spring-fall and fall-spring growth in grade 2 were based on the same proportion of total growth in each interval as was evidenced in the first grade. That is, the ratio of total spring-to-spring growth that took place from spring-to-fall obtained from explicit first grade norms was used to estimate second grade fall norms from the total second grade growth from spring-to-spring.[3] While the assumption of equal growth patterns in first and second grades may not strictly be valid, it may provide a more realistic (at least more conservative) guess as to the true growth curve than does linear interpolation. Comparisons of the estimated fall norms are made in Figure 3.

[3] The three explicit norms for first grade are:

Grade	Mean	Standard Deviation
0.7	31.3	12.2
1.2	35.6	13.7
1.7	59.4	18.4

The growth ratio used for the mean was: $\frac{35.6 - 31.3}{59.4 - 31.3} = 0.15$

The growth ratio used for the standard deviation was: $\frac{13.7 - 12.2}{18.4 - 12.2} = 0.24$

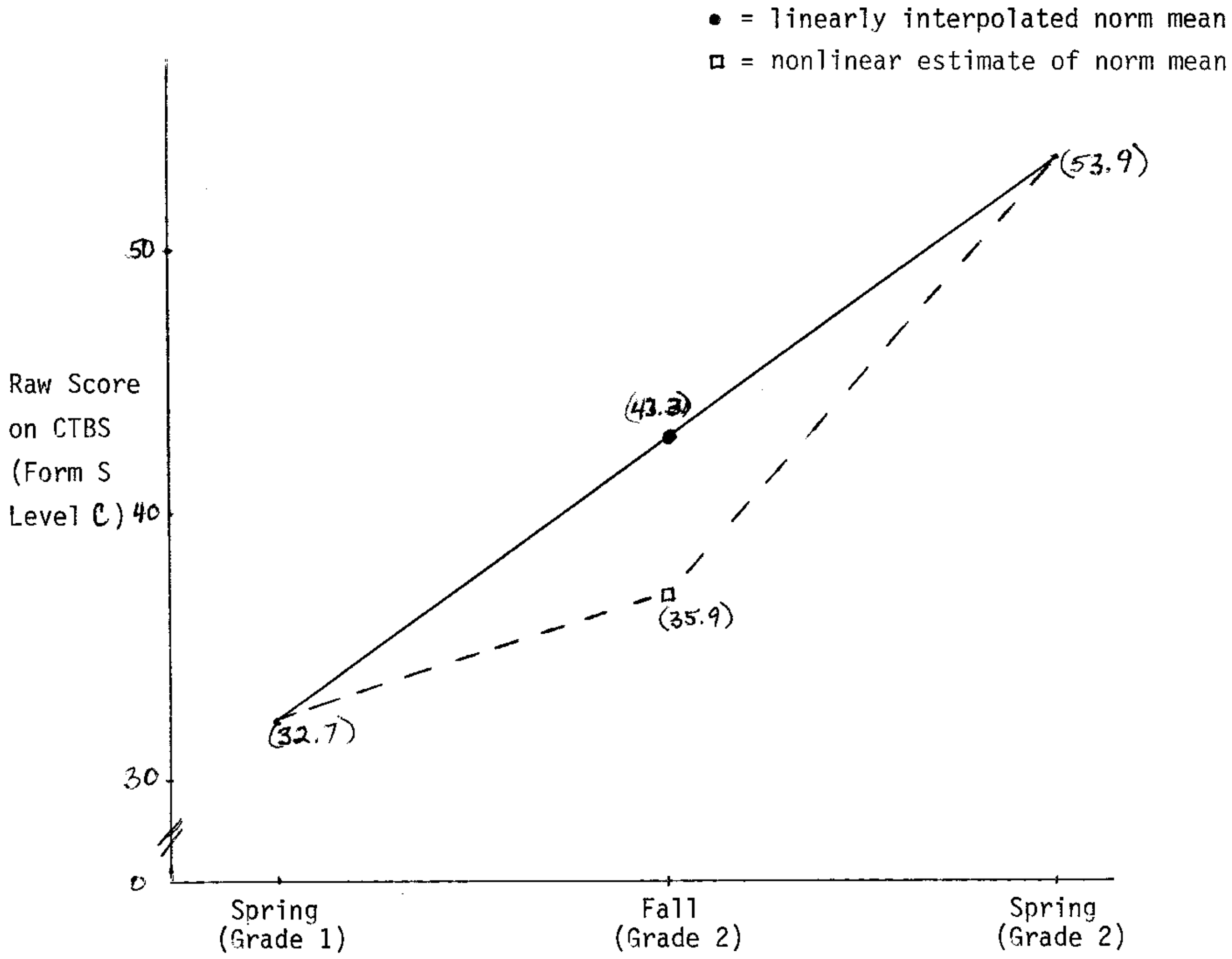


FIGURE 3 Published norm means for spring grade 1 and spring grade 2 (Form S Level C of CTBS) and two estimates of the fall norm mean are plotted. The solid line represents the single linear function used to adjust the norms in the exact linear procedure. The dashed line represents the two separate linear functions used to adjust norms in the exact-nonlinear procedure.

Whereas one of the three procedures (the "fixed-linear" procedure) used fall and spring norms as fixed values, the other two linearly adjusted the fall and spring norms according to the date of testing. The "exact-linear" procedure which also made use of linearly interpolated fall norms used a single linear function (the line from spring to spring norms) to adjust for testing date. Specifically, day to day growth was linearly estimated on the basis of a 300 day school year (30 days summer growth) and norms were adjusted in accordance to the number of days before or after November 1 or April 1 that testing took place. In the "exact-nonlinear" procedure which estimated fall norms on the basis of first grade growth ratios, two linear functions were used to adjust for testing date. For tests administered prior to November 1 an adjustment based on the 150 days of growth from previous spring to current fall was used. For tests administered after October 30 (including all posttests) the adjustment was based on linear growth from current fall to spring. Figure 3 shows the growth lines along which adjustments were made.

The actual transformations to standard score form were carried out as follows. Raw scores were subtracted from their appropriate norm means and divided by their norm standard deviations. The resulting value was then converted to a scale with mean of 50 and standard deviation of 10. Standard score gains were obtained by computing the difference between pre- and posttest standard scores. Using this scale the norm standard score is 50 and the norm standard gain is zero.

RESULT

Though testing dates vary considerably within each month, the three different types of standard scores have been averaged for all schools testing in a given month for the purposes of the display. This averaging artificially smooths the graphs and most likely reduces the actual day-by-day differences among the three transformation procedures. The transformed pretest scores and the standard score gains are tabulated by month of pretest in Table 1. They are also plotted in Figure 4. The reader should be aware that the plotted points for November and December are probably based on too few points to reflect a stable trend. Examination of the plots are limited here to the months of September and October.

We immediately see that the common fixed-linear procedure provides the lowest pretest transformed scores for early testing and that scores increase by month of pretest. The resulting gains obtained are the highest of all three procedures and exhibit a steady decline with month of pretesting. The pretest standard scores obtained using the exact-linear method are higher than those for the fixed-linear approach and do not exhibit the positive slope indicating the spurious advantage of early pretesting. Its resulting gains are less than the fixed-linear gains though they still exhibit a slight negative slope with month of pretest. The results for the exact-nonlinear procedure seem to approximate expected norm group growth. The pretest standard scores are centered around the norm of 50, and show a slight negative slope with date of testing. The corresponding gains are less than the norm gains of zero and appear to be relatively stable with pretest date.

Though we have only investigated the effects of early pretesting, the same types of effects (but opposite) would be expected with late posttesting.

(a) Pretest standard score means by transformation procedure, subject, and month of pretest for schools reporting the same form of CTBS fall and spring.

	FIXED-LINEAR		EXACT-LINEAR		EXACT-Nonlinear	
Month	Reading	Math	Reading	Math	Reading	Math
September (111,105)*	42.5	42.7	44.6	45.8	50.2	51.4
October (261,241)	43.3	43.3	43.6	43.0	48.5	47.7
November (9,11)	45.0	46.7	43.6	44.0	47.6	47.9
December (4,4)	52.7	48.4	49.7	45.4	52.8	48.3

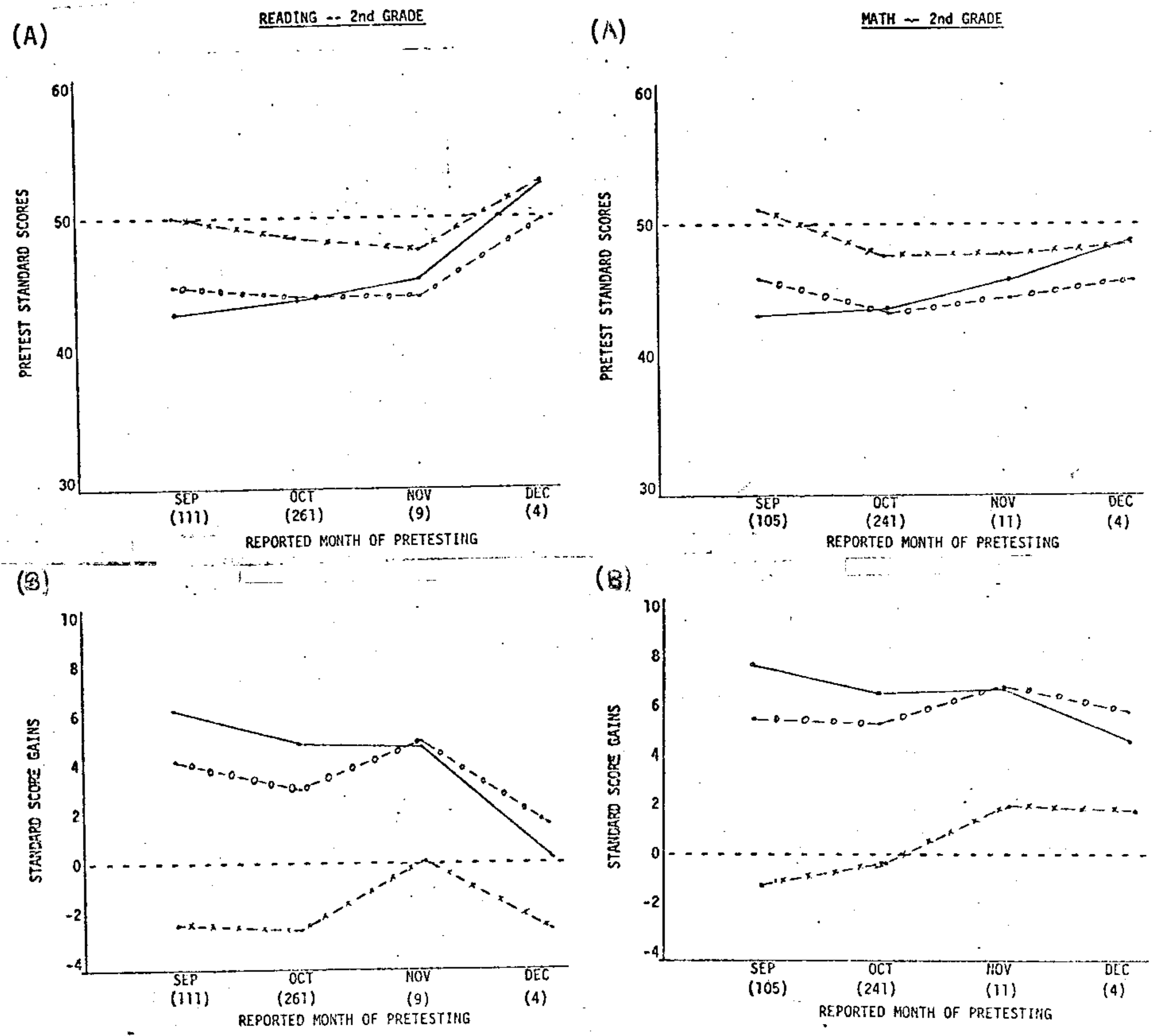
(b) Standard score gains by transformation procedure, subject, and month of pretest for schools reporting the same form of CTBS fall and spring.

	FIXED-LINEAR		EXACT-LINEAR		EXACT-Nonlinear	
Month	Reading	Math	Reading	Math	Reading	Math
September (111,105)	6.2	7.9	4.1	5.7	-2.6	-0.9
October (261,241)	4.9	6.6	3.0	5.3	-2.8	-0.4
November (9,11)	4.7	6.5	4.9	6.6	0.0	1.8
December (4,4)	0.1	4.1	1.5	5.4	-2.7	1.4

*The numbers in parentheses represent the number of schools upon which the mean is based for reading and math, respectively.

TABLE I Pretest standard score means and mean standard score gains for 1974-5 CTBS test results for three different transformation procedures.

Figure 4. Comparison of three methods of transformation for 1974-75 CTBS reading and mathematics scores in grade two (2). Pretest standard scores (A) and standard score gains (B) are presented by month of pretest.



*Number of schools pretesting in the given month(i.e., number of schools on which means are based).

LEGEND: ————————FIXED LINEAR INTERPOLATED NORMS
x-x-x-x-x-Exact Nonlinear Interpolated Norms
o-o-o-o-o-Exact Linear Interpolated Norms

The fact that posttest dates have not been held constant, and that schools pretesting early had a tendency to posttest earlier, results in gain score patterns which may not fully indicate the effects of differential testing dates. For instance, the slope of gains by month of pretest is still somewhat negative for the exact-linear method whereas we might expect it to be flat.

DISCUSSION

The results appear to support our earlier hypotheses. Furthermore, they provide clarification and empirical examples for Tallmadge and Horst's (1974) warnings concerning date-appropriate norms. Pretest standard scores and pre- to posttest gains <u>are</u> affected by the use of norms adjusted to the date of testing. In addition, the method used to estimate fall norms can drastically alter conclusions about academic growth.

It seems apparent that the use of linearly interpolated and fixed norms to transform raw scores can result in the overestimation of standard score gains. Using such a procedure, there is a clear advantage in pretesting early (and, presumably, posttesting late). Schools pretesting early tend to have lower pretest standard scores and higher resulting gain scores. Thus they are likely to show more academic growth relative to the norm group than are schools pretesting later. Furthermore, the overall level of apparent growth from all schools is likely to be inflated. Indeed, the results show that participants in the compensatory education programs at hand have exceeded expected norm group growth by over one-half of a standard deviation in standard score units.

By adjusting the norms to the date of testing prior to transformation the advantages of earlier pretesting are diminished. Referring to the lines connecting standard scores for schools pretesting in September and October we see that the "exact-linear" line does not have the positive slope that the "fixed-linear" line does. However, since more schools pretesting earlier also posttested earlier, this effect is not so obvious with respect to gains by month of pretesting, i.e., the slope is still negative. Presumedly, if month of posttest were to be held constant the advantage of early pretesting, in terms of gains, would also be diminished by using adjusted norms. Overall, though, the magnitude of the gains are smaller for the exact-linear procedure than the fixed-linear approach. That is, the artificial inflation in gains caused merely by early or late testing has largely been eliminated.

Estimating fall norms based on the "dogleg" shaped growth curve and adjusting norms with two different linear functions have the effect of drastically eliminating the advantages of early pretesting. Pretest scores transformed in this manner are much higher than when using linearly interpolated norms, and the resulting standard score gains are much lower. In fact, in our analyses these gains appear to be lower than the "expected" norm group gains. Admittedly using first grade growth patterns to set second grade fall norms leaves much to be desired. It was merely used here to provide a lower bound comparison of standard gains. Since one criterion for program eligibility was low achievement, we would expect to find that the pretest standard scores of the program participants were somewhat lower than the expect norm group mean. Using the exact-nonlinear approach though has resulted in standard scores that are equal to the norm mean. It appears that the use of the first grade growth ratio has resulted in an underestimation of the second grade fall norms. In addition, since slopes of pretest standard scores by month of pretest are not flat, we can conclude that either there are drastic differences in growth patterns between low and average achievers or that we have not obtained an accurate model (i.e., the "dogleg") of second grade growth. Still, however, the results from using this procedure to estimate fall norms and to transform raw scores make explicit some of the problems in using linear estimation techniques and underscore the importance for a viable model of academic growth.

With regard to possible inferences and policy decisions resulting from these procedures, we can see large differences. Looking at the results of the fixed-linear approach one might conclude that the program was highly effective since the academic growth of program participants apparently exceeds norm group growth by more than one-half of a standard deviation. This could likely be taken as justification for program expansion. The exact-linear method also results in apparent growth due to program impact that is greater than that expected in the norm group. However, its gains are less than those for the fixed-linear approach. Still, though, such gains would be taken as evidence for program effectiveness. Schools pretesting early or posttesting late would not necessarily exhibit greater growth and more effective use of program funds than schools pretesting late or posttesting early, though. If raw scores were transformed using the exact-nonlinear method the same program would not appear to be effective with respect to academic growth. All schools, regardless of testing dates would tend to exhibit lower than average growth. Such results may be taken as evidence of program ineffectiveness and may lead to program cutbacks. According to which transformation procedure used, then, different decisions about the program's future may result.

Though we have reported results only for the second grade and for testing during the 1974-1975 school year, these same analyses were conducted on third grade scores and on 1975-1976 test results for both grades. In all cases the same contrasts that were exhibited here resulted. The use of adjusted norms eliminates the spurious advantage of early pretesting and late posttesting, and the use of "curve-estimated" rather than linearly estimated fall norms results in higher pretest standard scores and lower pre- to posttest gains.

The implications of these findings are clear. Every possible effort should be made to have all schools administer a test for which both fall and spring norms have been explicitly established, and to have testing administered as close as possible to the standardization dates. Furthermore, the evidence clearly points to the use of norms that are adjusted to the date of testing. In view of the limitations of linear estimation, though, the best procedure for adjusting norms remains unclear. Finally, if fall norms must be estimated such estimation should be based on some theorized growth curve. Reliance on assumptions of linear growth can seriously affect evaluation conclusions and resulting policy decisions.

REFERENCES

Beggs, D. L., & Hieronymous, A. N. Uniformity of growth in the basic skills throughout the school year and during the summer. Journal of Educational Measurement, 5, 91-97, 1968.

California State Department of Education. Evaluation Report of ECE, ESEA Title I, and EDY, 1975-1976, Sacramento, California, 1977.

California State Department of Education. Evaluation Report of ECE, ESEA Title I, and EDY, 1974-1975, Sacramento, California, 1976.

Cronbach, L. J., & Furby, L. How we should measure change -- or should we? Psychological Bulletin, 74, 68-80, 1970.

David, J. L. & Pelavin, S. H. Research on the Effectiveness of Compensatory Education Programs: A Reanalysis of Data, Final report: SRI project URU-4425, Stanford Research Institute, Menlo Park, California, 1977.

DeVito, J. J., & Long, J. V. The effects of spring-spring vs. fall-spring testing upon the evaluation of compensatory education programs. Paper presented at the annual convention of the American Educational Research Association, New York City, April, 1977.

Keesling, J. W., & Burstein, L. An Audit Report on the Activities of the State Department of Education Related to the Early Childhood Education Program. Volume II, Evaluation of the California Early Childhood Education Program. Los Angeles: Center for the Study of Evaluation, University of California, 1977.

Tallmadge, G. K. & Horst, D. P. A Procedural Guide for Validating Achievement Gains in Educational Projects. RMC Report No. UR-240, RMC Research Corporation, Mountain View, California, 1974.