DOCUMENT RESUME

ED 220 492 TM 820 505

AUTHOR Bunch, Michael B.
TITLE Using Non-Normed Tests in Title I Evaluation.
PUB DATE Mar 82
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Compensatory Education; Correlation; Criterion Referenced Tests; Elementary Secondary Education; *Evaluation Methods; Models; *Program Evaluation; Scores; Test Format; *Testing Problems; *Test Norms; Test Use
IDENTIFIERS Elementary Secondary Education Act Title I; *Non Normed Tests; *RMC Models

ABSTRACT

Research evidence relating to the utility of the RMC evaluation models of compensatory education employing non-normed tests is examined. The history and evolution of five early models into the current norm-referenced model utilizing a non-normed test (Model A2), non-normed versions of a comparison model (Model B2), and the regression model (Model C2) are described. Technical considerations of minimum correlation, score overlap and score distribution for Model A2 and the estimation of population standard deviations on the non-normed test for Models B2 and C2 are discussed. Practical considerations include group size, grade level, type of score used and instrument sensitivity. Problems and proposed solutions are discussed. The preference for non-normed tests is shown to stem from a perceived discrepancy in the sensitivity of tests to instructional objectives, so the task of comparing alternative implementation strategies for various models is considered. The utility constraints of the Education Consolidation and Improvement Act of 1981 are discussed. In the light of practical alternatives presented and seemingly insurmountable problems in Models A2, B2 and C2, it is recommended they be abandoned. (Author/CM)

USING NON-NORMED TESTS IN TITLE I EVALUATION

Michael B. Bunch
Measurement Incorporated

Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, March, 1982.

USING NON-NORMED TESTS IN TITLE I EVALUATION

The U.S. Office of Education (now the U.S. Department of Education), pursuant to provisions of the Education Amendments of 1974 (Public Law 93-380), developed three evaluation models for education programs funded by Title I of the Elementary and the Secondary Education Act of 1965 (Public Law 89-10). Each of these models has two versions. In version one of each model, only normed referenced tests are used. In version two, both normed and non-normed tests are used. These models (there are six in all) as well as the supporting system (the Title I Evaluation and Reporting System or TIERS) are described in detailed in a user's guide (Tallmadge and Wood, 1976). Explicit reference to the evaluation models was made in the 1978 Education Amendments (Public Law 95-561), and subsequent regulations required the use of these models. Now, however, with passage of the Education Consolidation and Improvement Act of 1981 (Public Law 97-35; Title V, Subtitle D) the federal role in evaluation has become one of providing nonbinding guidelines.

In light of the changing role of the federal government in the evaluation of compensatory education programs, it seems appropriate to examine these models. The specific purpose of this paper is to highlight research evidence relating to the utility of the versions of the models employing non-normed tests. A related objective is to provide some guidance to those individuals at the national level who will establish the nonbinding guidelines and to the individuals at the state and local levels who will have to choose among several evaluation strategies.

This paper traces some of the historical developments of the various evaluation models currently used in Title I evaluation. In addition, some of the problems encountered in the application of these models are discussed

along with some of the solutions that have been proposed. Finally, recommendations relating to the relative appropriateness of the various models are given.

Historical Perspective

In the early stages of development of the current evaluation models, there were actually five models (U.S. Department of Health, Education and Welfare, undated). These models are briefly described as follows:

- Model 1 - Posttest Comparison with Matched Groups. "This model requires that children be paired in terms of pretest measures and that one member of each pair be randomly assigned to the treatment group and the other to the comparison group." (Ibid. p.49).
- Model 2 - Analysis of Covariance. Analysis of covariance provides an appropriate statistical adjustment to compensate for pretest score differences between groups if these differences were due to such chance factors as random sampling fluctuations. (Ibid. p.54)
- Model 3 - Special Regression Models. This was actually two regression models, one based on the regression projection model (Tallmadge and Horst, 1974). The other based on the regression discontinuity model (Campbell and Stanley, 1963).
- Model 4 - General Regression Model. This model is actually more similar to the analysis of covariance model. Essentially, several independent variables may be used to develop a regression equation to predict posttest scores.
- Model 5 - Normed Referenced Model. "Project children are compared to a norm group usually comprised of a nationally representative sample of children at the same grade level. The no-treatment expectation is that the project pupils will maintain, at posttesting, the same achievement status with respect to the norm group as they had at pretesting." (U.S. Dept. of HEW, p.72).

In the description of these models, there is no reference to the use of non-normed tests. It is apparent that the intent of the model developers was to use standardized tests only. However, upon completion of extensive field testing of the five evaluation models and discussions with state and local education officials in every state in the country, developers reduced the number of models to three, and added a non-normed version to each model

(Gamel, Tallmadge, Wood, and Binkley, 1975).

Thus, the Title I Evaluation and Reporting System that exists today is based primarily on a compromise between technical excellence and political reality. The posttest comparison with matched groups (Model 1) and analysis of covariance (Model 2) have been combined to form the comparison group model (Model B). The Campbell and Stanley version of the regression model (i.e. the regression discontinuity model) has been dropped in favor of the Tallmadge and Wood regression projection model and is currently known as the special regression model (Model C). The generalized regression model (Model 4) was dropped entirely, and the normed referenced model (Model 5) has survived as the norm referenced model (Model A). Furthermore, each model now allows for the use of non-normed tests. This allowance is clearly the result of input from state and local education officials.

Within each of the six evaluation strategies, all program effects are described in terms of Normal Curve Equivalents (NCEs). This metric consists of a standard score scale with a mean of 50 and a standard deviation of 21.06. The scale was constructed so that the NCE value and percentile value would be the same at 1, 50, and 99. Other aspects of NCEs are described by Tallmadge and Wood (1976).

In the non-normed versions of each model, score gains are estimated by linking a normed test to a non-normed test. In the version of the norm referenced model which utilizes a non-normed test (Model A2), the two tests are linked through an equipercentile equating procedure, and gains on the non-normed test are translated into estimated normed test gains. In the non-normed version of the comparison model (Model B2) and the regression model (Model C2), gains are expressed in terms of the hypothetical distribution of

scores on the non-normed tests. Estimation of population parameters, specifically standard deviations, becomes the key technical issue in the use of Models B2 and C2. While applications of Model A2 had been fairly common in many states, the use of Models B2 and C2 is extremely rare.

Problems Encountered

Problems encountered in the use of Models A2, B2, and C2 divide fairly neatly into two categories: practical and technical. Many of the practical problems described below are experienced in all three of the models. Some of the practical problems described below (for example, testing at or near the empirical norming date) are also observed with the normed versions of each of the models. The technical problems encountered, however, clearly divide the models into two groups: Model A2 and Models B2 and C2. Since the technical problems for Models B2 and C2 are so similar, these two models are grouped together throughout the remainder of this paper. Any discussion of problems or proposed solutions for either of these two models should be considered appropriate for the other. Discussions of proposed solutions for Model A2 should not be generalized to Models B2 and C2 unless specifically noted otherwise.

Before describing the problems encountered with Model A2, perhaps it would be helpful to look at the way in which Model A2 developers intended for it to be implemented. In Model A2, the following situation is typically found. An evaluator tests pre and post with a non-normed test (either a locally developed or a commercially available criterion referenced test) and administers a normed test either as a pretest or as a posttest. The exact steps to be carried out (assuming a normed pretest) are as follows:

1. Administer at pretest time non-normed and nationally normed tests, according to normative data points for the normed test.

2. Obtain the correlations between the normed and the non-normed test for the population. If the correlation is less than .60, use Model A1.

3. Determine median pretest raw score for the normed test.

4. Determine national percentile from pretest norms table which corresponds to median raw score, representing the expected no treatment effect.

5. Convert the no treatment percentile to an NCE, representing the expected no treatment effect.

6. Administer the identical non-normed test at posttest time, according to normative data points for the normed test (that is, at or near the empirical norming date of normed test).

7. Determine the median post test raw score for the non-normed tests.

8. Convert the median post test raw score to a pretest percentile (i.e. determine how many students scored below that point at pretest time).

9. From pretest norms, find the normed test raw score correpsonding to the percentile obtained in step eight.

10. From posttest norms, find the normed test percentile corresponding to the raw score obtained in step nine.

11. Convert this percentile to an NCE.

12. Subtract the results of step five from the results of step 11. This is the observed Title I effect.

This process is referred to as equipercentile equating at the median only. The same process may be applied, with some modifications, if the normed test is administered as a posttest. The technical problems associated with the implementation of Model A2 have to do with the correlation between the norm referenced test and non-normed test, the overlap of scores from pretest to posttest, and what are commonly referred to as floor and ceiling effects in either the non-normed test or the normed test. Other, practical problems may also arise. These practical problems involve group size, grade level,

instrument sensitivity, and type of score used (either raw score or mastery score on the non-normed test). The research relevant to each of these issues is presented below along with some of the proposed solutions.

Although the procedures described above for the implementation of Model A2 do not use the coefficient of correlation between the normed and non-normed tests at any point, a low correlation casts extreme doubt on the usefulness of the evaluation results. As noted previously, the models developers recommend a minimum correlation of .60 (Tallmadge and Wood, 1976). Even when the minimum correlation of .60 is obtained, the two tests share only 36% common observed variance. Several investigators have shown that even this minimum correlation of .60 may be very difficult to obtain under normal circumstances (cf., Storely, Rice, Harvey, and Crane, 1979; Bunch and Dixon, 1980; Kahn and Overton, 1980).

A somewhat more subtle technical problem has to do with the overlap of scores from pretest to posttest of the non-normed tests. Specifically, step eight of the implementation procedures requires that the median posttest raw score be converted to a percentile on the pretest score distribution. If the average student obtains a posttest raw score higher than that of the highest scoring student on the pretest, then Model A2 cannot be implemented. This situation is illustrated below in Figure 1.

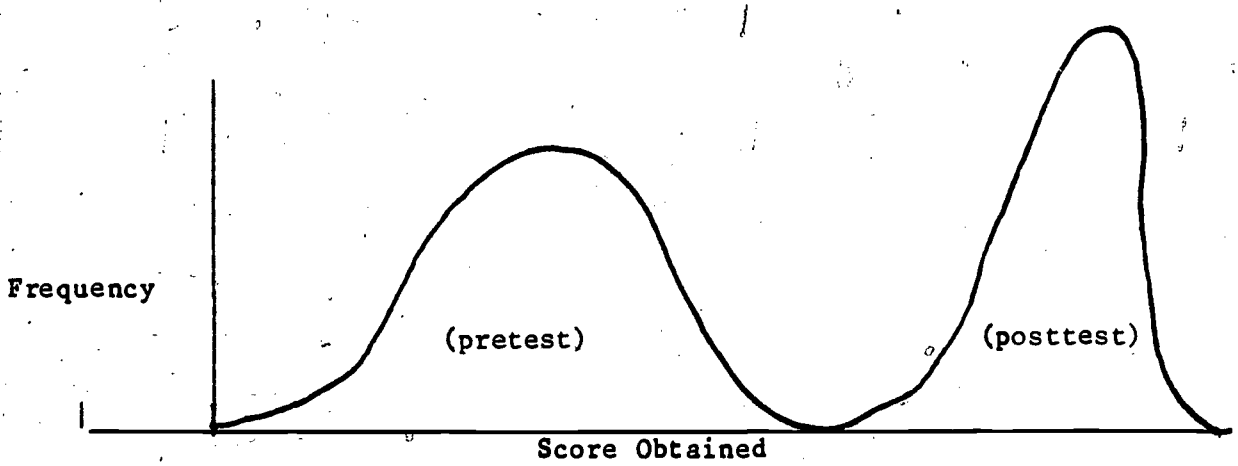

Figure 1. Hypothetical distribution of pretest and posttest scores on a non-normed test.

The problem of score overlap has been discussed in a monograph produced for the U.S. Department of Education by RMC Research Corporation (U.S Department of Education/RMC Research Corporation, 1981) as well as by Gamel (Note 1) and Bunch and Dixon (1980). This problem is closely related to the problem of choice of score (i.e., total raw score vs. number of objectives mastered). It can be shown, for example, that in choosing the objectives mastered indicator as the pretest and posttest score, the range at both pretest and posttest times will be extremely restricted. If, on the other hand, raw scores are used, there is more likely to be a spread of scores at both the pretest and at the posttest. However, even when raw scores are used, the overlap between pretest scores and posttest scores is likely to be minimal, if instruction was effective (cf., Popham, 1978).

One problem frequently found in all types of program evaluation is the problem of test floor or ceiling effects. Floor effects are those effects observed when the test administered to students is too difficult. Consequently, most students receive very low scores. Additionally when the test is a multiple choice test (as are nearly all normed tests), the average score may approach the score that might be obtained by chance guessing. The

mean score is thus too high (not too low as some might contend). Ceiling effects on the other hand, occur when a test is given that is too easy for the achievement or ability level of the student population to which it is administered. Since the upper limit of achievement is not tapped, mean scores are too low.

In the case of Model A2, floor and ceiling effects present a double problem. That is, not only does one have to worry about floor and ceiling effects on the norm referenced test but one must deal with these effects in the non-normed test as well. Specifically, when one uses only two tests as in Model A1 (a norm referenced test administered at pretest and at posttest time), there are nine possible combinations of floor, and ceiling effects. Of these nine combinations, only one will yield an appropriate estimate of the gain score. When three tests are used, the number of possible combinations of floor and ceiling effects is 27. Of these 27 possibilities, only one will yield an appropriate estimate of the gain.

Brummet and Masters (1980) presented data illustrating the problem that occurs when a ceiling effect is observed on the non-normed test at posttest time. Their data showed that under these conditions, Model A2 systematically underestimated the size of gain for the Title I program. Their sample focused on a single project in a school district where it was later observed that evaluations for other projects were seriously flawed with both floor and ceiling effects.

Crane, Prapuolenis, Rice, and Perlman (1981) used computer generated data to test the effects of various model violations on the outcomes of Model A2. One of the conditions considered was extremely positively or negatively skewed score distributions on the non-normed tests. Their skewed score distributions

correspond very closely to the distributions of scores observed when floor or ceiling effects occur. A major finding of the study by Crane et al. (1981) was "when Model A2 is applied with CRT data as negatively skewed as observed in Chicago Title I CRT data, all of the equating procedures examined will result in considerably biased NCE gain estimates." (p.4)

In 1978, the U.S. Office of Education/Office of Planning, Budget, and Evaluation requested the formation of a national committee to examine the norm referenced evaluation model (Model A). This committee contained a subcommittee to examine problems associated with Model A2. In October of that year, the subcommittee on Model A2 made the following recommendations regarding the implementation of Model A2:

(1) the TACs (Technical Assistance Centers) should provide for a hierarchy of strategies for analyzing non-normed test data that will allow LEAs of varying size and technical sophistication to select a particular strategy. A tenative hierarchy was proposed by the subcommittee:

a. Develop normalized test score distributions for the non-normed and normed tests at the local level.

b. Use a curvilinear analog for equating test across the entire score range as described in Angoff (1971).

c. Use a linear analog for equating tests as described in Angoff (1971).

d. Use the current A2 strategy, however, if this alternative is chosen, it is recommended that the TACs make efforts to assist LEAs in: (1) the development or selection of a non-normed test, and (2) the selection of an appropriate level of a normed test in order to avoid the possible effects of score range distribution (Hansen, 1978, pg. 24).

The previously cited study by Crane et al. (1981) was a direct response to the Model A2 Subcommittee Report. Crane et al. investigated the effects of several different levels of NRT-CRT correlation, levels of sample size, and levels of treatment effect on the amount of bias introduced into NCE gains estimates obtained by the four different equating procedures described above. Data in the Crane et al. study were computer simulated.

Among the findings presented by Crane et al. were several analyses comparing the four equating methods across treatment conditions, sample size, and correlation level. In virtually every instance, all of the procedures proposed by Hansen's group produced smaller errors than those produced by the Model A2 procedures described earlier in this paper as the standard Model A2 procedures.

Bunch and Dixon (1980) carried the Crane et al. analyses a bit further by allowing for both forward equating and backward equating in Model A2.[1] Bunch and Dixon examined seven methods of estimating gains. These seven methods are presented below:

Method 1 - Standard Model A2, predicting posttest from pretest

Method 2 - Standard Model A2, predicting pretest from posttest

Method 3 - Regression Method, predicting posttest from pretest

Method 4 - Regression Method, predicting pretest from posttest

Method 5 - Linear Equating, predicting posttest from pretest

Method 6 - Linear Equating, predicting pretest from posttest

Method 7 -Standard Model A1; the standard of comparison for methods 1 through 6.

The equations for Methods 1 through 6 are described in detail by Bunch and Dixon (1980).

Bunch and Dixon examined two separate data bases in the comparison of these equating methods. The first data base came from a Virginia school division where TIERS Models A1 and A2 had been simultaneously implemented.

[1]The reader will recall that the equating in Model A2 may be either at pretest or at posttest. When one equates at pretest, one predicts the posttest scores. When one equates at posttest, one predicts pretest scores.

This data base contained some of the problems discussed previously (for example low correlation and lack of score overlap).

The Virginia data base consisted of pretest and posttest scores for 138 students. Each student had taken one of seven levels of a locally developed criterion referenced test in the fall of 1978 and again in the spring of 1979. Additionally, an appropriate level of the SRA Achievement Battery (Science Research Associates, 1974), was administered at both testings. Each student thus had four scores; normed pretest, normed posttest, non-normed pretest, and non-normed posttest.

Because seven levels of the CRT were administered, there were fourteen CRT-NRT correlations (seven pre, seven post). Of these fourteen, only one reached the minimum acceptable level of .60. Thus, Methods 1 and 2 became technically infeasible, and for all practical purposes, Methods 3 through 5 became practically unfeasible. However, where correlation between normed and non-normed tests was relatively high, estimates based on correlations (Methods 3 and 4) were generally fairly accurate. The regression methods appeared to work better overall than linear equating methods or the standard Model A2 when the results from Model A1 were used as the criterion.

The second data base was that used by Pellegrini, Horwitz, and Long (1979). In that study, two standardized tests had been given to groups of third graders (N = 64), fourth graders (N = 55), fifth graders (N = 58), and sixth graders (N = 81). Each student had been administered an appropriate level of the California Achievement Test (CAT) Form C and the Comprehensive Test of Basic Skills (CTBS) Form S on two separate occasions. Thus, each student in the sample had four standardized test scores.

These data were chosen because either of the two normed tests could be treated as a non-normed test, and Methods 1 through 7 could be applied accordingly . With the CAT-CTBS data, there was no clearly superior method of equating. In some instances, the forward regression method seemed to work best, while in other instances the standard Model A2 predicting posttest from pretest seemed to work better.

One finding of the second study was the discrepancies between gain estimates produced by the California Achievement Test and the Comprehensive Test of Basic Skills. Since both of these tests are norm referenced, this may sound like a problem of test selection for Model A1. However, the pattern of discrepancies was similar to the pattern of discrepancies in gains when one compares norm referenced tests and criterion referenced tests.

In most instances, the norm referenced test will produce a greater dispersion of scores than will the criterion referenced or non-normed tests. In this specific instance, because large raw score gains on the CAT are required to produce modest NCE gains, such gains translate into extremely large gains on the CTBS where relatively small gains are capable of producing NCE gains. Similarly, a small score gain on a criterion referenced test may translate into a relatively large NCE gain on the standardized or norm referenced test. One is forced to wonder whether or not this particular feature of Model A2 has been its primary selling point over the past several years.

Some of the problems previously discussed may have been exacerbated by the relatively small size of the sample involved. Gamel (Note 1) analyzed several studies and concluded that a sample size of 300 would be adequate for implementation of Model A2. Gammel went on to suggest that Model A2 is a good model for aggregation at the state or federal level. However, particularly at the local level, a sample size of 300 seems unlikely.

With respect to grade level, Game1 (Note 1) found that gains become more stable at the higher grades. It should also be pointed out, however, that gains also tend to decrease at the higher grades (See Note 2). Thus, Model A2 may appear to be a suitable model at the higher grade levels with fairly large sample sizes. Again, one runs into the practical problem that there are fewer and fewer students involved in Title I as they progress through elementary school and into junior high school and high school (See Note 2). In other words, where Model A2 appears to work best, there appear to be the fewest numbers of students on whom it may be used.

Related to the technical problem of low NRT-CRT correlation is the practical problem of differential instrument sensitivity. This problem was discussed by Fish (1979) and could easily account for some of the results found by Linn (1979). This practical problem seems to be at the very heart of the decision to use Model A2. Because Model A2 requires the administration of three tests rather than two, there must be some practical advantage to using it. Most users of Model A2 contend that the criterion referenced test is more sensitive to their instructional program than their standardized norm referenced test (c.f. Note 1). In these instances, it is very frequently the case that the standardized norm referenced test is one used by the entire school district for general testing purposes. This test may or may not be appropriate to the Title I objectives and practices. The non-normed test, on the other hand, is quite frequently developed specifically for the Title I program or by Title I staff in conjunction with other school staff for specific instructional objectives taught by Title I teachers and regular classroom teachers alike.

That tests do not all measure exactly the same thing is amply pointed out by Porter, Schmidt, Floden, and Freeman (1978). They show that there are major differences in content among commonly used standardized norm referenced tests. How much greater then is the content difference between nationally produced, standardized, norm referenced tests and locally produced, criterion referenced tests?

Indeed, if the scenario just mentioned is fairly widespread, then it would seem that Model A2 be in use in a very biased sample of school districts across the nation. That is, where the locally used norm referenced test appears to Title I evaluators to be sensitive to the Title I and regular classroom objectives, Model A1 will be used. Where the district-wide test does not appear to be content valid to the Title I evaluator, Model A2 is more likely to be used.

However, Title I program effects are expressed in terms of gains on the norm referenced tests. Thus, gains are being explained on a fairly insensitive measure. The more insensitive the measure the more likely it is that Model A2 will be used. Thus, the more insensitive the norm referenced test, the lower the correlation between normed tests and non normed tests is likely to be. Model A2 would therefore appear valid only in those cases where it is not greatly needed.

The problems and associated research for Models B2 and C2 are less extensive than those of Model A2. All reports to date have focused on the issue of estimating population standard deviation on a non-normed test, the main underlying concept of the two models. Long, Horwitz, and DeVito (1978) showed that the standard procedure for estimating the population value of the standard deviation on the non-normed test (σ^2_{nn}) was systematically biased.

Specifically, this estimate was systematically high when the wrong level of the normed test was used and systematically low when the non-normed test scorerange was restricted. Since Title I test scores are quite likely to be very restricted, it seems reasonable to expect that $\sigma^2_{\eta\eta}$ would usually be underestimated.

Bunch (1979) proposed a solution to the estimation of $\sigma^2_{\eta\eta}$. This solution was based on previous work by Cronbach (1971) and relies on sample correlations and sample standard deviations. This proposed solution was subsequently criticized by Pellegrini, Long, and Horwitz (1979). Pellegrini et al. argued that the solution proposed by Bunch also systematically underestimated $\sigma^2_{\eta\eta}$ and thus systematically overestimated the Title I effect. Bunch (1979) developed an alternate formula, again based on work by Cronbach, but this time using sample correlations between normed and non-normed tests corrected for attenuation. This alternative produced more reasonable estimates of the population of standard deviation on the non-normed tests and thus better estimates of the Title I effect. One serendipitous finding of the Bunch study was that scale scores are far superior to raw scores for this type of estimation (Tallmadge and Wood (1976) had recommended using raw scores). In every instance, whether the original formula for the estimation of the population standard deviation was used or either of the alternatives were used, the estimate was consistently more accurate when scale scores, rather than raw scores, were used.

There has been virtually no investigation of Models B2 and C2 since 1979. Furthermore, there appears to be have been little if any reaction to the proposed solutions offered at that time. However, as noted earlier, use of Models B2 and C2 is quite rare or perhaps non-existent.

Conclusions and Recommendations

The utility of Models A2, B2, and C2 has been examined from the perspectives of technical adequacy and practicality. The technical issues have focused on minimum correlation, score overlap, and score distribution for Model A2 and the estimation of population standard deviations on the non-normed test for Models B2 and C2. The practical considerations have included group size, grade level, type of score used, and instrument sensitivity. It is on this last issue that the distinction between practical and technical concerns becomes blurred.

In the final analysis the utility of any model is relative. Since the driving force behind the selection of Models A2, B2, and C2 seems to be a preference for non-normed tests, and since a great deal of this preference seems to stem from a perceived discrepancy in the sensitivity of tests to instructional objectives, one is faced with a task not simply of comparing models but of comparing alternate implementation strategies for various models. For example, during the period under study (roughly 1974 to 1981) most major test publishers have re-normed their standardized achievement tests. As older versions of these tests are removed from the market place, local evaluators have been faced with the task of selecting different tests or perhaps purchasing the newer version of the old tests. Having examined the instructional sensitivity of the old tests, many evaluators have lobbied strongly in their districts for the selection of more instructionally sensitive norm referenced tests. Indeed, the recommendation to do so has been at the heart of much of the technical assistance provided by the Title I evaluation Technical Assistance Centers.

As more and more school districts adopt norm referenced tests that are highly instructionally sensitive, the motivation for using non-normed tests in

Title I evaluation will diminish. This phenomenon has been observed in many school districts served by the author throughout HEW (now ED) Region III (Pennsylvania, Delaware, Maryland, District of Columbia, Virginia, and West Virginia). Indeed, the number of applications of Model A2 in Virginia, for example, has diminished by approximately 75 per cent (from 22 down to 5) over the last three years. Most evaluators who have dropped Model A2 cited practical considerations.

Perhaps the most damaging blow to the non-normed models has been the emphasis in the 1978 amendments and the 1981 legislation on sustained effects. The Education Consolidation and Improvement Act of 1981 stresses evaluation that covers a period of time of at least one calendar year. This emphasis presents severe constraints for the non-normed models. These models have typically been used in fall-spring testing programs. A typical case would involve the administration of the normed and non-normed pretest in the fall and a non-normed posttest in the spring for Model A2. With the emphasis on twelve-month evaluations, many districts have adopted a spring-spring testing schedule. Given this testing schedule, it is necessary to administer a norm referenced test each spring. To administer the non-normed test and then estimate scores on the subsequent spring normed test seems somewhat ludicrous when the actual scores on that test are available. Thus, even where Model A2 might seem feasible, one is forced to weigh its feasibility against the imposition of additional testing. In this situation, once-a-year testing has won out more often than not.

Less testing which serves the same purpose as more testing would obviously seem to have greater utility. At the same time, if two tests measure instructional objectives equally well and one has the added advantage of national norms, the test with national norms would seem to have a greater utility.

When Models A2, B2 and C2 were first introduced, there was little emphasis on once a year testing and a great deal of making do with whatever norm referenced test happened to be available. Thus, the availability of more appropriate norm referenced tests and the emphasis upon full year evaluations have worked together to undermine the relative utility of non-normed models.

The measurement of change or growth will forever be fraught with problems (cf., Harris, 1963). In every instance, our best estimate of the amount of growth produced by a particular program or project is simply that, an estimate. With non-normed models, we end up with estimates of estimates. In light of practical alternatives to models A2, B2, and C2, and given the seemingly insurmountable problems inherent in these models, it seems totally appropriate to advocate abandoning them as we approach a new era in the evaluation of compensatory education programs.

REFERENCE NOTES

1. Nona N. Gamel - The adequacy of Model A2. Draft report prepared for the Department of Health, Education & Welfare, U.S. Office of Education/Office of Evaluation and Dissemination by RMC Research Corporation, Mountain View, California, April 1980.

2. RMC Research Corporation - Preliminary reports to Office of Planning and Evaluation, U.S. Department of Education, 1982.

REFERENCES

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Brumet, M. & Masters, B. Floor and ceiling effects with Model A1 and A2. In Lewis, J.S., (moderator) On the uses of criterion referenced tests in Title I Evaluation: A time and place for everything. Symposium presented at the annual meeting of the Eastern Educational Research Association, Norfolk, Virginia, March 1980.

Bunch, M.B. Linking normed and non-normed tests for evaluation of Title I programs. Paper presented at the annual meeting of the Eastern Educational Research Association, Kiawah Island, South Carolina, February 1979.

Bunch, M.B., and Dixon, R., Linking Normed and Non-normed Tests in TIERS Model A2: Sometimes you can't get there from here. In Lewis, J.S. (moderator), On the uses of criterion referenced tests in Title I Evaluation: A time and place for everything. Symposium presented at the annual meeting of the Eastern Educational Research Association, Norfolk, Virginia, March, 1980.

Campbell, D.T., & Stanley, J.C. Experimental and quasi-experimental design for research on teaching. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Crane, L.R., Prapuolenis, P.G., Rice, W.K., & Perlman, C. The effect of different equating methods on Title I evaluation Model A2 NCE gain estimates: Final Report (USDE Contract #300-79-0485). Evanston, Illinois; Educational Testing Service, 1981.

Fish, O.W. An analysis of the evaluation data when ESEA, Title I evaluation models A1 and A2 are empirically field tested simultaneously. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April 1979.

Gamel, N., Tallmadge, G.K., Wood, C.T., & Binkley, J.L. State ESEA Title I Reports: Review and Analysis of past reports, and development of a model reporting system and format (report prepared for the U.S. Department of Health, Education and Welfare). Mountain View, California; RMC Research Corporation, 1975.

Hansen, J.B. Report of the Committee to Examine Issues Related to the Use of the Norm Referenced Model for Title I Evaluation, Northwest Regional Educational Laboratory, October, 1978.

Kahn, L., & Overton, W. Model A2 from a practical point of view: Which way to go? In Lewis, J.S. (moderator), On the uses of criterion referenced tests in Title I evaluation: A time and place for everything. Symposium presented at the annual meeting of the Eastern Educational Research Association, Norfolk, Virginia, March 1980.

Linn, R.L., Validity of inferences based the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, 1, 15-22.

Long, J., Horwitz, S., and DeVito, P. An empirical investigation of the ESEA Title I Evaluation System's proposed variance estimation procedures for use with criterion referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Pellegrini, A.D., Long, J.V., and Horwitz, S. An empirical investigation of the ESEA Title I Evaluation System's proposed variance estimation procedures and the proposed alternative for estimation of variances in Title I evaluation models for use with criterion referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, April, 1979.

Popham, W.J. Criterion referenced measurement. Englewood Cliffs, N.J.: Prentice Hall, 1978.

Porter, A.C., Schmidt, W.H., Floden, R.E., & Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15, 529-540.

Storlie, T.R., Rice, W., Harvey, P., & Crane, L. An empirical comparison of Title I NCE gains with Model A1 and A2. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, April, 1979.

Tallmadge, G.K., and Wood, C.T. User's Guide: ESEA Title I Evaluation and Reporting System. Mountain View, California: RMC Research Corporation, 1976.

U.S. Department of Health, Education & Welfare. A practical guide to measuring project impact on student achievement. Washington, D.C.; U.S. Department of Health, Education & Welfare/Office of Education, undated.

U.S. Department of Education/RMC Research Corporation. Evaluator's references: Title I Evaluation and Reporting System, Volume II. Washington, D.C.: U.S. Government Office, 1981.