DOCUMENT RESUME

ED 156 696 TM 007 121

AUTHOR Linn, Robert L.
TITLE The Validity of Inferences Based on the Proposed Title I Evaluation Models.
PUB DATE Mar 78
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; Compensatory Education Programs; Control Groups; Data Analysis; Elementary Secondary Education; Equated Scores; Evaluation Methods; Mathematical Models; *Models; *Norm Referenced Tests; Norms; Program Effectiveness; *Program Evaluation; Research Design; Research Problems; Sampling; Standardized Tests; *Validity
IDENTIFIERS *Elementary Secondary Education Act Title I

ABSTRACT

The three RMC models endorsed by the U.S. Office of Education for the evaluation of Elementary and Secondary Education Act Title I programs are based on narrowly conceived approaches to evaluation--the measurement of cognitive achievement gains. Each model requires the comparison of observed student performance with an estimate of what level of performance would have been achieved without the Title I project. The models differ in the way the non-participation estimate is made. Under Model A the estimate is based on test norms, assuming equal mean pretest and mean posttest percentile ranks. Model B requires the use of an experimental design including a control group. Model C is described as a regression-discontinuity analysis. Only Model A appears to be feasible for most Title I programs. However, data are presented to challenge the plausibility of the assumption that non-participation would result in the percentile rank of the posttest mean being equal to the percentile rank of the pretest mean; and additional data are presented to challenge the assumption that the use of an equating test might ameliorate this problem. It is proposed that educational evaluation is too immature for fixation on a small set of models. Flexibility is needed. (Author/CTM)

The Validity of Inferences Based on the Proposed Title I Evaluation Models*

Robert L. Linn

University of Illinois at Urbana-Champaign

* Paper presented as part of a symposium entitled "The RMC Models for Evaluation of Title I, ESEA -- A Look at Validity and Utility" at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 27-31, 1978.

Printed in U.S.A.

The U.S. Office of Education has recommended three models for the evaluation of Title I projects. These models were developed by the RMC Research Corporation and are commonly referred to as the "RMC Models". As stated by the developers "The focus of all the models is to obtain as clear and unambiguous an answer as possible to the question 'How much more did pupils learn by participating in the Title I project than they would have learned without it?'" (Tallmadge & Wood, 1976, p. 2).

It has been claimed that the models "... are generally acknowledged to be technically sound..." and "[i]f properly implemented, each will yield valid, comparable, and interpretable results" (Tallmadge & Wood, 1976, p. 2). These are strong claims. With the USOE endorsement of the models and the amount of use that is apt to be made of the models in the evaluation of federally - funded education projects, these claims deserve close scrutiny. The purpose of this symposium is to provide some of the scrutiny, criticism and debate that is needed.

The focus of my presentation is on the issue of validity, particularly the internal validity of the RMC models. The bulk of my remarks will be directed at one of the three models. I will emphasize Model A, the norm-referenced model, because it is the least demanding of the three models from the user's perspective and is apt to be the most feasible to implement. Also, it is my understanding that it is the one that is currently receiving the greatest use.

A Brief Description of the RMC Models

A detailed description of the RMC Models is not possible within the time constraints of this presentation. Thus, my description will be

rather cursory. Those interested in a detailed description should see the report by Tallmadge and Wood (1976) entitled "User's Guide: ESEA Title I Evaluation and Reporting System" as well as several other publications prepared by staff of the RMC Corporation (e.g., Horst, Tallmadge & Wood, 1975; Tallmadge & Horst, 1976).

All three RMC models are based on narrowly conceived approaches to evaluation. Or at least they are limited to a single "... aspect of project evaluation, measuring cognitive achievement gains" (Horst, Tallmadge & Wood, 1975, p. 1). No attention is given to contextual or process variables. Nor, is any attention given to non-cognitive outcome variables. Even within the domain of cognitive achievement, little attention is given to questions about how achievement should be measured.

As previously stated, the purpose of each model is to answer the question "...:'How much more did pupils learn by participating in the Title I project that they would have learned without it?'" (Tallmadge & Wood, 1976, p. 2). In attempting to answer this question each model would require the comparison of the observed performance of students on an achievement test administered folllowing a period of participation in a Title I project with "... an estimate of what that performance would have been without the Title I project" (p. 2). The three models differ in the way this no-participation estimate is made. But, the main emphasis of all three models is on the generation of this estimate of what the achievement would have been for a group of Title I participants in the absence of the Title I project. For short this estimate is referred to as the "no-treatment expectation" but this label should not be interpreted literally since the absence of a Title I project would not, in fact, be anything like a "no-treatment condition".

Model A

Under Model A the no-treatment expectation is obtained from normative data. The underlying assumption on which the expectation is based is that, in the absence of Title I, the group mean would be at the same percentile at the time of the posttest as it was the time of the pretest. If the percentile rank at the time of the posttest is greater than the percentile rank at the time of the pretest then a positive estimate of the project impact would be obtained. Obviously the percentile ranks must be determined from normative data appropriate to the particular time of testing.

For purposes of statistical analysis and aggregation of data to the state and national level, the observed and expected no-treatment posttest scores would be expressed in normal-curve equivalents (NCE's). The NCE's are simply normalized standard scores with a mean of 50 and a standard deviation of 21.06. They range from 1.0, which is the NCE corresponding to a percentile of 1.0, to 99.0, which is the NCE corresponding to a percentile of 99.0. In comparison, to percentiles, NCE's are more spread out toward the middle of the distribution and less spread out at the extremes with equal numerical values at 1, 50 and 99.

As stated the model is based on the use of a standardized norm-referenced achievement test with national norms available at dates closely corresponding to the pretest and posttest administration dates. Variations in the basic model allow for the use of a locally normed test or even a non-normed test. The latter variation is referred to as Model A2 and since I will have more to say about this variation I'll provide a brief description of it at this point.

It is possible for project staff to select a test or to develop a test for use in Model A2. Thus, a test that is considered particularily appropriate for evaluating a project, may be used despite the absence of normative data. The desired test would be administered as above, pre and post. At pretest time, however, it would ne necessary to also administer a nationally normed test. No-treatment expectations would be based on the results of an equipercentile equating of the normed and non-normed results at the time of the pretest. Normative data on the normed test at dates corresponding to the pretest and posttest administrations would be used to define the no-treatment expectation and the equipercentile equating would be used to provide to the link back to the observed results on the non-normed test of interest.

The remaining two models may be described even more concisely than I have described Model A, partially because they are more familiar paradigms and partially because I will have less to say about them in my critique.

Model B

Model B is readily described as it comes directly from notions of basic experimental design. It is called the control group design. The idealized, albeit seldom realized, application of model B would involve random assignment of children to "treatment" and control groups. Both groups would be administered a pretest and posttest. The tests could be normed (model B1) or non-normed (B2), but if the latter a normed test would also be needed to convert to the NCEs required for aggregation purposes.

Model C

The last model is the "special regression design". Where feasible it is considered preferable to model A but less desirable than model B.

The special requirement of model C is that assignment to a Title I project be based exclusively on the pretest. All children below a cutoff would be assigned to the project and all those above the cutoff would not participate in the Title I project but would be retained as a comparison group. In other words, Model C is what Campbell & Stanley (1963) described as "regression-discontinuity analysis".

As with the other models, model C may be used with normed or non-normed tests. With non-normed tests a normed test would also need to be administered to obtain NCEs, however.

Critique

There are many questions that could be raised about the validity of these models for purposes of evaluating Title I projects. Obviously, learning is not the only outcome of concern and even if it were there are serious questions that could and should be asked about the sensitivity of the broad survey achievement tests that are given preference under the models. They were designed for other purposes and may contain few items of particular relevance to the activities of a given project, not nearly enough to be considered adequate as the sole indicator of project impact.

In addition to concerns about the narrow definition of student outcomes, concerns could be raised about the lack of attention to educational process or the context in which the project is implemented. Without such information an evaluation is at best incomplete, and may be a waste of time, or even misleading.

In the brief time available, I must focus on only a few of the many questions of validity that might be raised. But, this is not intended to imply that issues such as the ones just mentioned are unimportant.

Quite the contrary, they are of vital importance.

I shall focus on just two questions concerning Model A. These are:

1. Is there an adequate basis for using a constant percentile as the no-treatment expectation?

2. Is it reasonable to use norms for one test to establish the expected no-treatment performance level for another test?

Before turning to these questions a few brief comments about the validity and feasibility of the other two models and why I choose to focus on Model A are in order.

Models B and C -

If viewed as research designs the three RMC models are easily ranked in terms of their relative internal validity. In its idealized form Model B is a classic experimental deisgn and ranks highest in terms of internal validity. Model A ranks third with Model C somewhere in between. This ranking agrees with the stated order of preference provided by the developers of the models.

While the order of models is clear in a research context it is not so clear within the context of evaluating Title I projects, which is where the models are meant to be applied. The researcher's paradigm can seldom be imposed upon a school system. Random assignment of children causes not only operational problems but is apt to be opposed on philosophical grounds. A positive answer to the impact question is generally assumed in advance of the evaluation and it is considered unethical to deny the advantage to some of the most needy solely for purposes of the evaluation.

Even if random assignment were possible the researcher's ability to control the treatments is not apt to be available to the evaluator. Thus, the

distinctions between the "treatment" and "control" groups are apt to be blurred. Thus, the front-runner, Model B, cannot be applied in idealized form. At best, it is only approximated in practice and the internal validity of Model B is apt to be seriously compromised.

Without random assignment there will always be questions about the comparability of the "treatment" and "control" groups and about how adjustments should be made for pre-existing differences. There is a relatively large literature on these issues, which I shall not attempt to review here. (See for example, Campbell & Erlebacher, 1970; Cronbach, Rogosa, Floden & Price, note 1; Lord, 1967; Rubin, 1974). Suffice it to say, there are no fool-proof solutions. Pre-existing differences are a substantial threat to internal validity.

The statistical adjustments that are recommended for use with Model B rest on strong assumptions. If these assumptions do not hold then the adjustments can yield seriously biased estimates. (See Linn, note 2 Linn & Werts, 1977).

In addition to the technical problems, Model B ranks low on the dimensions of cost and feasibility. It would seem problematic enough for a local project to be able to identify an adequate non-participant comparison group. But the problem is further exascerbated by the need to get this comparison group, once identified, to participate in the data collection for the evaluation.

If Model B cannot be used then, from a research design perspective the next most rigorous model is Model C. The latter is only applicable, however, under very specialized circumstances. It could be used only for a project where participants were identified <u>solely</u> on the basis of their pretest scores. Hence the design is relevant only to a particular and

probably small subset of projects. In those specialized circumstances Model C is potentially useful. But, its validity will depend upon strong assumptions especially on the assumption of linearity.

The remaining model is clearly the most feasible. The norm-referenced model may be implemented with relative ease. It does not require the identification of a control group as in Model B or the specialized circumstances needed for Model C. Given that one of the three models is to be used, Model A is apt to be the most frequent choice. But, is it apt to be sufficiently valid to be worth using? It is with this more general question in mind that I am finally ready to turn to a more detailed consideration of the two questions that I asked earlier regarding Model A.

Question 1

Is there an adequate basis for using a constant percentile as the no-treatment expectation?

Defining the no-treatment expectation to be equal to the posttest score with the same percentile rank that was achieved on the pretest rests on an assumption that in the absence of an intervention, the group would maintain the same relative standing over time. A stringent test of this assumption for target groups of interest is not possible since it would require observations over time in the absence of special interventions. That is no more feasible than the creation of randomly selected control groups.

Although a stringent test of the constant percentile rank assumption is not possible, it can be evaluated in a more limited sense by asking whether the current state of affairs is for groups to maintain roughly constant percentile ranks or standard score positions. If that was found

to generally be the case, then it could be used as the basis for determining if new programs alter the status quo.

Van Hove, Coleman, Rubben, and Karweit (note 3) summarized achievement tests results for two grade levels in New York, Los Angeles, Chicago, Philadelphia, Detroit and Baltimore. Tests results for these cities were available in grade 6 and either grade 3 or grade 4 on one of three standardized achievement test batteries. Schools were categorized by percent of minority students in the school. Unweighted average percentile ranks were then obtained for groups of schools categorized by percent minority. As a global summary the average percetile ranks over parts of the same test battery were calculated and reported for schools where students were nearly-all minority group members and for schools with nearly-all majority group students.

I have converted the global results reported by Van Hove, et al to NCE Scores (Linn, note 2). The resulting NCE's are reported in Table 1 for "earlier grade", which is either grade 3 or grade 4, and for "later grade", which is grade 6 in all instances. The difference between the NCE of the later and earlier grade is also reported for each category of schools and each city and test combination. With the exception of city A, the later grade NCE is lower than the earlier grade NCE for the nearly all-minority schools in all cities. The later grade NCE's are also lower than the earlier grade NE's in 5 of the 7 instances for nearly all-majority schools. The decline for the all-majority schools is generally less than that for the all-minority schools, however.

Insert Table 1 about here

The data on which the results in Table 1 are based leave a number of things to be desired for our purposes. For instance, they are cross-sectional rather than longitudinal, the level of the test is not constant across grades, and averaging of percentile ranks across parts of a battery may conceal interesting trends on parts of a battery. Nonetheless, they are sufficient to raise some doubts about the universal applicability of the constant NCE score as the expected no-treatment effect. The unweighted average of the difference in NCE's for the nearly all-minority schools in Table 1 is -2.8 and ranges from -7.7 to +4.4. It at least seems debatable that a zero NCE difference is the right no-treatment effect expectation.

The option of using either local or national norms is also called into question by data such as those in Table 1. In cities where the test results tend to fall behind the national norms over increasing grades the local norms are sure to seem preferrable from the perspective of the project director.

Kaskowitz and Norwood (1977) have reported results of several analyses that are relevant to the question of the adequacy of the constant percentile expectation. They compared the pretest and posttest percentiles for a longitudinal norming sample (Beck, note 4) with cross-sectional norms on the Metropolitan Achievement Test. When converted the NCE units the posttest NCE was slightly higher than the corresponding pretest NCE at most grades for the longitudinal norms group as a whole. For grades 2 thru 8 the difference on the Total Reading score ranged from -0.2 to +2.9 NCE units. The corresponding figures for math were -1.2 to +5.5 (Linn, note 2). Though statistically significant due to the large sample at each grade, these differences may be considered relatively small. They do provide evidence that cross-sectional norms may not yield the same information as longitudinal norms would.

Furthermore, a systematic error of even 2 NCE units may be of concern if the size of the effects to be detected is small.

Possibly more important than the above Kaskowitz and Norwood results were their findings for students with extremely low pretest scores and their findings for minority students. They found that the constant percentile assumption led to an expected posttest score that was too low for students with extremely low pretest scores. When the constant percentile expectation was used for minority students and for students in the comparison group of the Follow Through data reported by Apt Corporation, however, the so called no-treatment expectation was too high. Kaskowitz and Norwood concluded that the "[u]se of the norms based on the standardization group will lead to an expected posttest score that will be too high for students ordinarily in compensatory education programs, expecially minority students who have pretest scores that are not extremely low" (1977, p. 55).

These results underscore the tenuous nature of the key assumption on which model A is based. That is the assumption that in the absence of an effect due to special treatment, the achievement level of the group would maintain a constant NCE as defined by the norm group. The tenuous nature of this assumption was acknowledged as a weakness of Model A by Horst, Tallmadge and Wood (1975) who noted that "empirical support for this assumption is most plausible "... when the norm group is like the treatment group ..." (p. 72). By defninition however, the samples used to develop national norms on standardized tests are not like the specially defined subpopulations to which compensatory education programs are directed.

Thus, I would suggest that the answer to my first question is no. There is not an adequate basis for the constant percentile assumption on which Model A is based. Even if the question were answered in the affirmative, however, the procedures for the use of Model A2 would be problematic. Thus, I will turn to my second question.

Question 2

Is it reasonable to use norms for one test to establish the expected no-treatment performance level for another test?

The flexibility to select and use a normed test will be welcome by many who view the existing non-normed tests as insensitive to educational effects. The requirement of administering a normed test in addition to the desired one might be considered a nuisance, but one worth tolerating in order to be able to use a specially constructed test or an available criterion- or domain-referenced test that was judged to be more sensitive to the specific effects of the project.

As in Model A1, the no-treatment expectation for Model A2 is based on the assumption that the project participatns would maintain the same percentile rank over time in the absence of the equating of the normed and non-normed tests.

One of the rules of implementation for Model A2 is that the "... normed and non-normed tests should measure approximately the same ability..." (Tallmadge & Wood, 1976, p. 45). The tests must be highly correlated. The minimum correlation mentioned by Tallmadge & Wood is .6. This standard is much to lenient.

The tests equated in the Anchor Test Study (Loret, et al, 1974) had intercorrelations substantially higher than .6, yet even that equating would result in noticeable errors if used as in Model A2. To illustrate

this, ATS results for the Sequential Tests of Educational Progress (STEP) and the California Achievement Tests (CAT) were used. Suppose that the STEP Reading Comprehension Tests was administered at grade 4 and the median raw score was 10. This raw score corresponds to a percentile rank of 28 and an NCE of 37.7. If the STEP was to be administered in grade 5 for the posttest then the no-treatment expectation would be a median raw score of 11.8 which is the raw score corresponding to an NCE of 37.7 in the 5th grade norms.

Now suppose that STEP was a non-normed test. To use it in model A2 the CAT reading comprehension subtest is administered at the pretest time in grade 4 along with the STEP reading comprehension subject. The procedure for using Model A2 would then be based on the equating of the STEP and CAT grade 4 raw scores. The CAT norms for grade 4 and 5 would be used to define the no-treatment expectation. I have simulated this process using the ATS grade 4 equating results and the ATS norms for the CAT in grade 4 and 5. The procedure and the data source for each step is outlined in Table 2.

Insert Table 2 about here

The results of the simulation are shown in the bottom half of Table 3. As indicated by the double-headed arrow in Table 3, this may be compared to the corresponding value of 11.8 that would have been obtained under Model A1 (see the top half of Table 3).

Insert Table 3 about here

The STEP raw score no-treatment expectation of 11.8 is the value that would be obtained using Model A1. The value of 13 is the result of the

simulated application of Model A2. In terms of NCE units these two expected no treatment outcomes are 37.7 and 41.3 respectively. A difference of 3.6 NCE units may seem large to some but acceptably small to others. I will not argue that point but merely note that even under equating conditions much better than can generally be expected for Model A2 applications, systematic errors may be introduced simply due to the equating.

The correlation of the STEP and CAT reading comprehension subtests in grade 4 was .76 for the ATS sample. Correlations even lower than this may be expected when non-normed tests are correlated with normed tests in model A2. Jaeger (note 5), for example, noted that Athey and O'Reilly had found predictive validities of a criterion-referenced test with the CAT Total Reading Subscore ranging from .37 to .69. Grandy, Werts, and Schabacker (1977) reported correlations of .73 and .70 between the Georgia Criterion Referenced Tests in reading and the Total Reading Score on the Iowa Tests of Basic Skills at grades 4 and 8 respectively. The corresponding figures for math were .71 and .76.

How disparate might the results of model A2 be from those based on an actual norming of the non-normed tests with a correlation of only about .7 between the normed and non-normed tests?. I know of no data to directly answer this question. I have used data reported in the CAT manuals (Tiego & Clark, 1970, 1972), however, to simulate the type of results that might realistically be obtained.

The correlation between the Math Computation Subtest and the Math Concepts and Problem Subtest on Form A, Level 2 of the CAT was reported to be .70 for the grade 2.6 normative group (Tiego & Clark, 1972). The correlation is at the desired level for the simulation. The two subtests

are also of interest for my purposes because, although they are both part of the Total Math Subscore, they might be expected to be differentially sensitive to instruction between grades 2.6 and 3.6. Indeed the CAT item data shows large increases in proportion passing on the computation items between grades 2.6 and 3.6, particularly on the multiplication items. The corresponding increases for the concepts and problems items are smaller.

For purposes of my simulation, I assumed that the Concepts and Problems Subtest was the normed test and the Computation Subtest was the non-normed test. My imaginary project director presumably opted for the non-normed test because it contained those items on which large gains were expected as the result of the instructional program. For percentiles of 15 through 50 in steps of 5 percentile points the grade 2.6 Computation Scores were used to obtain two no-treatment expected NCE scores at grade 3.6. First the actual grade 3.6 Computation norms were used. In other words the percentiles were simply converted to NCE scores. Second, the posttest expected NCE's under no-treatment effect assumptions were obtained as they would be in model A2. That is, the grade 2.6 Computation and Concepts-Problems scores were equated, the grade 3.6 Concepts and Problems NCE was then used to identify the corresponding raw score, which was in turn used in the grade 2.6 equating to obtain the equivalent Computation raw score. Finally, the Computation raw score was converted to an NCE using the grade 3.6 Computation Subtest Norms. Normally this last step would be impossible since the test in Model A2 is non-normed.

The resulting expected no-treatment effect NCE scores are shown in Table 4. Also shown is the difference between actual NCE and the "equated" NCE that would result from Model A2. These differences are of the order of

a third to a half of a norm group standard deviation. Differences of this magnitude surely would be considered too large to tolerate by any standards.

Insert Table 4 about here

I conclude that the answer to my second question is no. The norms on one test may be quite inappropriate for establishing a no-treatment expectation for another by means of an equating.

Concluding Remarks

The RMC models are intended to deal with a very difficult problem. The models were selected because they were considered "... feasible to implement in actual school settings" (Tallmadge & Wood, 1976, p. 2). The models are also expected to yield "... valid, comparable, and interpretable results" (Tallmadge & Wood, 1976, p. 2). There is of course, a tension between the goal of feasibility, which includes considerations of cost, level of disruption, and skill requirements for conducting the evaluation, and the goal of obtaining sound results.

Model A is high on the fesability dimensions but low in terms of its ability to provide "valid, comparable and interpretable results". It shares some of the notable features of its automotive namesake. It is very simple. It requires neither exotic statistical techniques nor expensive and typically infeasible control groups. On the other hand, I doubt that Model A can live up to its automotive namesake's reputation for dependability.

Models B and C also suffer from serious limitations due to compromises needed to make them feasible and due to the strong assumptions required when they are used in practical settings. If these threee models don't provide what is needed then what would? I will not propose another model. Indeed, I think that the enterprise of educational evaluation is far to immature for

a fixation on a small set of models. Flexibility is needed to allow for a variety of measurement procedures and approaches to evaluation. Such flexibility complicates the problem of aggregating information to the state or national level. But, it is far better to live with that complication than to force everything into a mold. The results of applications of the narrowly restricted RMC Models cannot be expected to provide much information that will be useful for improving the education of the children for whom the programs are intended.

Reference Notes

1. Cronbach, L. J., Rogosa, D. R., Floden, R. E. & Price, G. G. Analysis of covariance in nonrandomized experiments: Parameters affecting bias. Occasional Paper, Stanford Evaluation Consortium, Stanford, California: Stanford University, 1977.

2. Linn, R. L. Evaluation of Title I via the RMC Models: A Critical Review. Paper presented at the CSE Invitational Conference on Measurement and Methodology, Center for the Study of Evaluation, UCLA, Los Angeles, January 4-5, 1978.

3. Van Hove, E., Coleman, J. S., Rabben, K. & Karweit, N. School performance: New York, Los Angeles, Chicago, Philadelphia, Detroit, Baltimore. Unpublished manuscript, Baltimore, MD, October, 1970.

4. Beck, M. D. Development of empirical "growth expectancies" for the Metropolitan Achievement Tests, paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.: March, 1975.

5. Jaeger, R. M. On combining achievement test data through NCE scaled scores. Draft Report prepared for Research Triangle Institute Research Triangle Park, NC, (USOE Contract No. 300-76-0095), undated.

References

Campbell, D. T. & Erlebacher, A. Reply to the replies. In J. Hellmuth (Ed.), Compensatory education -- a national debate, Vol. 3, disadvantaged child. New York: Brunner Mazel, Inc., 1970, pp. 221-225.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook for Research on Teaching.

Grandy, J., Werts, C. & Schabacker, M. Equating of ITBS and Georgia CRT Reading and Mathematics Tests in the eighth and fourth grades. Princeton, NJ: Educational Testing Service, September, 1977.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement. Number 1 in a series of monographs on evaluation in educaiton. Washington, D.C.: U. S. Department of Health, Education and Welfare, 1975.

Kaskowitz, D. H. & Norwood, C. R. A study of the norm-referenced procedure for evaluating project effectiveness as applied in the evaluation of project information packages. Research Memorandum, Menlo Park, CA: Stanford Research Institute, January, 1977.

Linn, R. L. & Werts, C. E. Analysis implications of the choice of a structural model in the nonequivalent control group design, Psychological Bulletin, 1977, 84, 229-234.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.

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

Rubin, D. B. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology, 1974, 66, 688-701.

Tallmadge, G. K. & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Number 2 in a series of monographs on evaluation in education. Washington, D.C.: U.S. Department of Health, Education and Welfare, 1976.

Tiegs, E. W. & Clark, W. W. Bulletin of Technical data number 2: California Achievement Tests. 1970 Edition. Monterey, CA: CTB/McGraw Hill, 1972.

Tiegs, E. W. & Clark, W. W. Examiner's Manual California Achievement Tests 1970 Edition Level II, Form A. Monterey, CA: CTB/McGraw Hill, 1970.

Table 1

NCE Scores at Two Grades for Nearly All Minority Schools and Nearly All-Majority Schools in Several Cities[1]

		Nearly All Minority			Nearly All Majority		
City	Test[2]	Earlier Grade	Later Grade	Diff.	Earlier Grade	Later Grade	Diff.
A	1	30.7	35.1	4.4	41.9	45.8	3.9
B	1	33.0	29.9	-3.1	44.7	43.6	-1.1
C	1	29.9	26.3	-3.6	44.7	45.2	.5
D	1	33.0	25.3	-7.7	51.6	51.1	- .5
D	2	34.4	33.7	- .7	57.0	54.2	-2.8
E	2	40.1	33.7	-6.4	55.3	50.5	-4.8
F	3	23.0	20.4	-2.6	50.0	46.3	-3.7

[1] Based on average percentile ranks over parts of the same test battery reported by Van Hove, Coleman, Rabben and Karweit, 1970.

[2] The tests are (1) Iowa Tests of Basic Skills, (2) Metropolitan Achievement Tests, and (3) Stanford Achievement Tests.

Table 2

Procedure for Simulating Model A2 Results Using The Anchor Test Study Data

Step	Procedure	Data Source
1	Obtain CAT grade 4 mean raw score equivalent to STEP grade 4 mean raw score.	ATS grade 4 equating table
2	Convert CAT grade 4 raw score to percentile rank.	ATS grade 4 norms table
3	Convert CAT average percentile rank to no-treatment expected grade 5 CAT raw score.	ATS grade 5 norms table
4	Obtain STEP no-treatment expected grade 5 raw score.	ATS grade 4 equating table

Table 3

Hypothetical No-Treatment Expectations For the STEP Reading Comprehension Subtest Based on Models A1 and A2*

Test	Type of Score	Grade 4 Mean	Grade 5 No-Treatment Expectation
Model A1			
STEP	Raw	10.0	11.8
	NCE	37.7	37.7
Model A2			
STEP	Raw	10.0	13.0
CAT	Raw	17.0	21.0
	NCE	37.7	37.7

* The NCE's and the raw score equating of STEP and CAT are based on Anchor Test Study results (Loret, et al., 1974).

Table 4

Expected No-Treatment Effect Posttest NCE Scores Based On Actual Computation Subtest Nroms and Based on an Equipercentile Equating with the Concepts and Problems Subtest*

	Expected NCE		
Percentile	Actual Norms	"Equated" Results	Difference: Actual Minus "Equated"
15	28.2	21.8	6.4
20	32.3	24.8	7.5
25	35.8	27.5	8.3
30	39.0	29.6	9.4
35	41.9	32.3	9.6
40	44.7	34.7	10.0
45	47.4	36.5	10.9
50	50.0	38.1	11.9

*Based on Tiegs & Clark (1970, 1972). The subtests are parts of the California Achievement Tests.