DOCUMENT RESUME

ED 323 260 TM 015 506

AUTHOR McNeil, Keith
TITLE The One Group Posttest Evaluation Model.
PUB DATE 25 Jan 90
NOTE 13p.; Paper presented at the Annual Meeting of the Southwestern Educational Research Association (Austin, TX, January 25-27, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Compensatory Education; Content Analysis; *Curriculum Evaluation; *Educational Objectives; *Elementary School Students; *Evaluation Methods; Grade 2; Grade 3; Pretests Posttests; Primary Education; *Program Evaluation; School Districts
IDENTIFIERS Education Consolidation Improvement Act Chapter 1; *One Group Posttest Evaluation Model

ABSTRACT

Often there is a desire to evaluate a program, but there is no comparable comparison group available. This paper focuses on an evaluation model that can be used when there is no comparison group and when there is no pretest. The method--Model A--used by the Chapter 1 compensatory education program is described. Model A, which uses a pretest-posttest approach, is currently applied in most of the local educational agencies in the country. The one group posttest only design is advocated, which requires content specialists to identify which objectives on the posttest were included in the compensatory curriculum versus those included only in the regular curriculum. The compensatory students should perform better on compensatory objectives to which they were exposed in both the regular and compensatory programs than on regular curriculum objectives to which they were exposed only in the regular curriculum. Data from 2 successive years of an evaluation of a Chapter 1 program in Dallas (Texas) were analyzed. Data on 20 items were obtained from over 2,000 students each year at each grade level. The results were mixed, with third-graders performing significantly better on the district's criterion-referenced test items included in the compensatory curriculum, and second-graders performing better on items not included in the curriculum. When data for the two grades were combined, the results were in the expected direction. Results for the second year suggest that either the grade 2 compensatory curriculum or the implementation of that curriculum should be reviewed. Four analytical methods are outlined. Two data tables and eight figures are included. (TJH)

THE ONE GROUP POSTTEST EVALUATION MODEL

Keith McNeil
New Mexico State University

The One Group Posttest Evaluation Model

Keith McNeil
New Mexico State University
January 25,1990

Problem: Often there is a desire to evaluate a program, but there is no comparable comparison group available. One way to solve the problem is to look at the gain from pretest to posttest. (The Chapter 1 compensatory education program uses this design, calling it Model A.) (Horst, Tallmadge, and Wood,1975). But if it is also the case that the pretest is not available, then the program may have to remain unevaluated. The present paper provides an evaluation model that can be used when these two conditions exist--when there is no comparison group and when there is no pretest.

Model A of Chapter 1: A digression to the discussion of Model A is necessary at this time because that model is currently used in most of the LEAs in the country. Two major assumptions of Model A are usually not tenable in actual implementation. First, the pretest is often used to select students into the Chapter 1 program, thus allowing the regression effect to inflate the resulting "Chapter 1 effect." Second, the assumption that the regular program is of average effectiveness (the equipercentile assumption) is often not a valid assumption. Since Chapter 1 eligible students cannot be deprived of Chapter 1 services, a particular LEA cannot know how their Chapter 1 students would perform as a result of the regular curriculum only. See Figure 1 for a schematic of this assumption, and the top of Figure 2 for three possible Model A results with an effective Chapter 1 program, and the bottom of Figure 2 for three possible Model A results with an ineffective Chapter 1 program. As can be seen in Figure 2, Model A can often result in an incorrect conclusion, especially when one realizes that very few LEAs implement curricula of average effectiveness (for that LEA).

Procedures: The one group posttest only design avoids these two assumptions and as well can be utilized to evaluate a compensatory program when there is no comparable comparison group and when pretest data do not exist (Ryan, 1980). The design requires content specialists to identify which objectives on the posttest were included in the compensatory curriculum (the C objectives), and which objectives were included only in the regular curriculum (the R objectives). Figure 3 provides a schematic representation of a 20 item test with the R and C designations. The compensatory students should perform better on those C objectives to which they were exposed in both the regular and the compensatory program (the double dosing effect), than on those R objectives that they were exposed to only in the regular curriculum.

Paper Presented at the Annual Meeting of the
Southwestern Educational Research Association
January 25-27, 1990, Austin,Texas

Analysis I: One could compare the percent correct on the items measuring the two groups of objectives. The analysis would be a simple t-test of the difference between two groups--one group being the C items and the other group being the R items, as indicated in Figure 3, producing a result as in Figure 4.

Analysis II: It is possible that the items measuring the one group of objectives are of different difficulty than the items that are measuring the other group of objectives. The solution to this potential dilemma is to statistically equate the difficulty of the items by covarying the inherent difficulty of the items. One could use the difficulty information from either: 1)the norming sample, 2) the non-compensatory students in the same school, 3) the results from the non-compensatory students in the same school in previous years, or 4) the results from one or more LEAs using the similar curriculum and similar in demographics. Since the difficulty information is used only as a covariate, the adequacy of the information is not too crucial. That is, these additional groups are only providing information as to the difficulty of items on the posttest and the groups are not being used as comparison groups. The analysis would be a covariance analysis, covarying the difficulty of the items. The covariate is in the last column in Figure 3, and would produce a result as in Figure 5.

Analysis III: If one is concerned that the two lines in Figure 5 might not be parallel, then that assumption could be tested by allowing the two lines to interact, as in Figure 6. If indeed the lines were not parallel, then the evaluation would be providing valuable information to the curriculum people. The analysis would be a linear interaction between the difficulty of the item and the type of item.

Analysis IV: If one is concerned about the assumption of straight lines, then that assumption could be tested by allowing the lines to be curved, as in Figure 7. If indeed the lines were curved, then the evaluation would be providing valuable information to the curriculum people. The analysis would be a curvilinear interaction between the difficulty of the item and the type of item.

Interpretations: If one performed analysis I, then the mean difference between the two groups of items would be reported. Analysis II would result in the difference between the two lines being reported. Analysis III would call for the reporting of the difference at selected points along the interacting lines of best fit, whereas analysis IV would call for depicting the two curves of best fit. (See any general linear models text, such as McNeil, Kelly, and McNeil (1975) for statistical and reporting procedures.)

Data collected over a period of years at either the school level or at the LEA level could result in patterns such as those in Figure 8. Notice that all interpretations are strictly with regard to the Chapter 1 program and are irrespective of the effectiveness of the regular program. If Model A had been used with this data, different (and erroneous) conclusions would have been

obtained. If the regular program in each of these six LEAs was effective, each of these Chapter 1 programs might be considered to be effective (with the possible exception of LEA #6). On the other hand, if the regular program was not effective, these programs might be considered to be not effective (with the possible exceptions of LEA #1, LEA #2, and LEA #4).

Special concerns: The design rests heavily on the accuracy of the curriculum specialists being able to identify those objectives that were included in the two curricula. The task can be made a little easier by using a criterion-referenced test that has been designed to measure the regular curriculum. In such a case, the content people only have to identify those objectives that are in the compensatory curriculum.

In most school systems there is the additional assumption that the teachers actually taught the curriculum (and that the students listened to and learned from the curricula). The extent to which these assumptions are tenable causes problems for Model A as well, but only reduces the likelihood of obtaining significant results in favor of the compensatory program in the one group posttest only design.

An applied example: Data from two successive years of an evaluation of a Chapter 1 program in Dallas, Texas will now be presented. The district's criterion-referenced test (STEELS) that matches closely the state's essential elements was routinely administered as a posttest to both Chapter 1 and non-Chapter 1 students. The identification of which objectives were in the Chapter 1 "A Priori" compensatory curriculum was accomplished easily by the curriculum specialists. The inherent difficulty level of the items was determined from the non-compensatory students in the district as that data was readily available.

Results: Table one contains the results for the first year of implementing the evaluation design(McNeil, Berry, and Metze, 1988). When considering all three grades together, A Priori students did significantly better on items taught in the A Priori program than on items not taught in the program.When the results were viewed at each grade level, the results were always in favor of the A Priori program, but significance was obtained only at grade 1. The small number of items (the unit of analysis) at each grade level hampered the attainment of significance.

Table 2 contains the results for the second year of implementation of the new evaluation model (McNeil, Jones, Berry, Edoghotu, and Kane, 1989). In this year grade 1 students did not take the STEELS, so data was available for only grade 2 and 3. The results were mixed, with third-grade students performing significantly better on the items included in the compensatory curriculum and second-grade students performing better on the items not included in the curriculum. When the two grades were combined, the results were in the expected direction, but only approached significance.The results for the second year suggest that either the grade two compensatory curriculum or the implementation of that curriculum ought to be reviewed. It should be emphasized that these

results on 20 items were actually obtained from over 2000 students each year at each grade level. Although process evaluations did check on the quality of teacher implementation of the curriculum, student data was not eliminated if teachers did not implement the curriculum well.

Potential problems: Since this is a new design, one might wonder about whether or not there might be some problems in implementing the design. Although the one implementation discussed above resulted in no problems, several potential problems might be considered.

Calculations. As with any new evaluation model, ease in implementation is a reasonable concern. Analysis I is a straight-forward computation. Analyses II, III, and IV require an evaluator who understand covariance, interaction , and curvilinear interaction, respectively. For those who understand these concepts, the interpretive value of these analyses far outweigh the additional calculation burden. Existing computer packages such as SAS and SPSS can easily perform the calculations.

Aggregation of data. State and Federal evaluators want the data to be collapsible across LEAs. If the data are transformed to logits, a fairly straight-forward procedure, the results should be aggregatable.

Interpretation of results. The interpretation of results will have to rely on usage over time, as did the NCE metric when it was first introduced. It should be clear that the item level interpretations provide insights into curriculum, inservice, and teaching modifications that are not available with the current Chapter 1 evaluation models.

Determination of which curriculum items are in. This determination probably needs to be made by content specialists, rather than by evaluators. The task can be difficult and time consuming. On the other hand, one might argue that the content specialists should know both the regular and compensatory curricula well enough so that the task would not be that difficult, as was the case in the one application. In addition, such determinations are usually made when an LEA makes a test adoption decision. (One added benefit of this design is that the test adoption decision is less crucial for the compensatory program. Those items that are not in an LEA's curriculum or in the compensatory curriculum can be omitted from the analysis, which is not possible in the Model A analysis.)

Teacher implementation of curriculum. If the Chapter 1 teachers do not implement the Chapter 1 program as expected, then the analysis will wrongly accuse the Chapter 1 program of being not effective. Observation of Chapter 1 teachers could avoid this conclusion.

Only low difficulty items in the curriculum. A Chapter 1 curriculum might focus on low-level objectives, but most tests are designed such that each objective is measured by items of varying difficulty. If indeed the Chapter 1 curriculum is measured only by items of low difficulty, then analysis I will lead to an incorrect conclusion, but analyses II, III, and IV will still be applicable.

Testing out of level. Many compensatory students take a lower level test, as recommended by the developers of the Chapter 1 evaluation models (Roberts, 1981). Since the same kind of curriculum fit determinations can be made with an out of level test as with an on level test, testing out of level would not cause a problem with the new evaluation model.

Summary: An evaluator may on occasion be confronted with the need to produce an evaluation of a compensatory program when there is no available comparison group and when no pretest data is available. The design discussed in thins paper provides a tool for obtaining an evaluation under such constraining circumstances, without sacrificing any evaluation principals.

The design is particularly valuable for two reasons. First, few, if any, evaluators ever find a perfect comparison group in the real world. In this design, the students serve as their control. Second, if program gains are evaluated over a school year, which they usually are, it may be inappropriate to use the same test for both pretest and posttest. It may be very difficult to identify a test which adequately measures the objectives desired at the posttest and which can be administered at pretest.

NOTE: I would like to thank Joe Ryan for initially discussing this design, and Napoleon Mitchell, Gail Smith, Wayne Murray, William Denton, George Powell, James English, and David Vines for forcing me to have a better conceptualization of the design. I especially want to thank Barbara Mathews, Jane Seibert, and Rosie Ramirez for identifying the items and helping me chart the unknown.

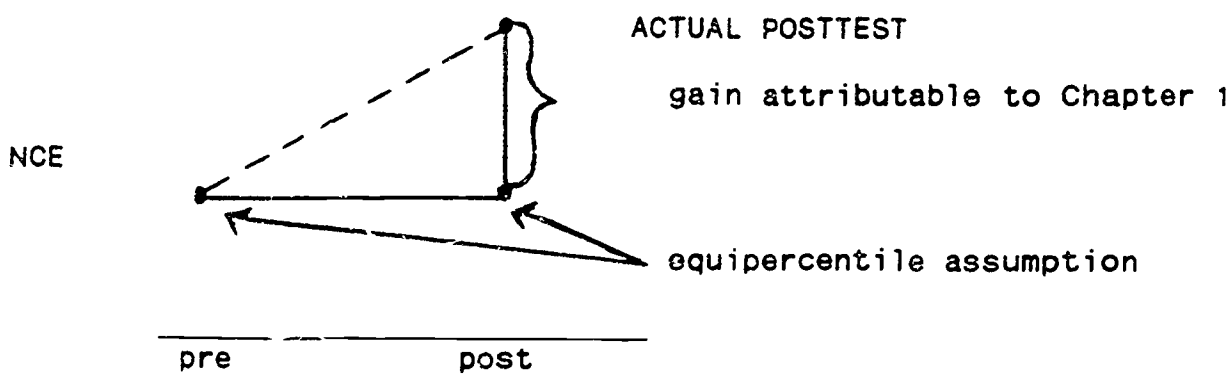
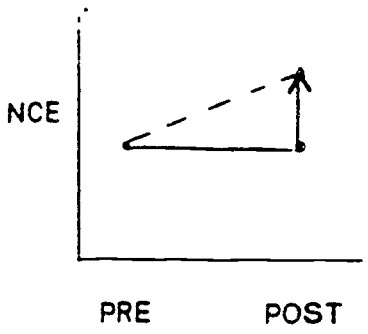


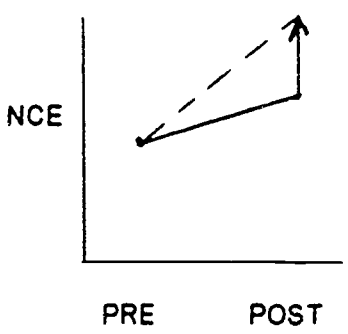
Figure 1. How the NCE gain is calculated in Model A.



Regular curriculum:
average effect

Chapter 1:
effective

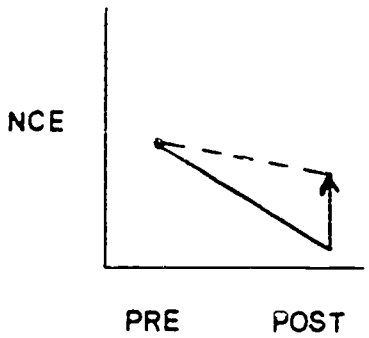
Model A:
effective



Regular curriculum:
above ave. effect

Chapter 1:
effective

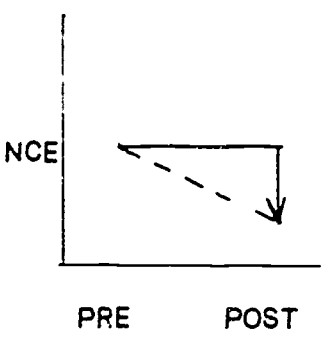
Model A:
above ave. effect



Regular curriculum:
not effective

Chapter 1:
effective

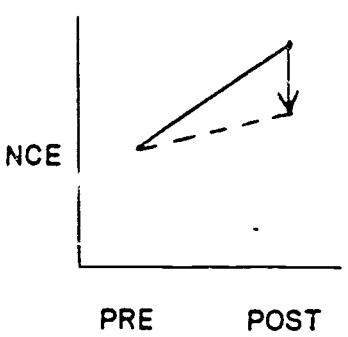
Model A:
not effective



Regular curriculum:
average effect

Chapter 1:
not effective

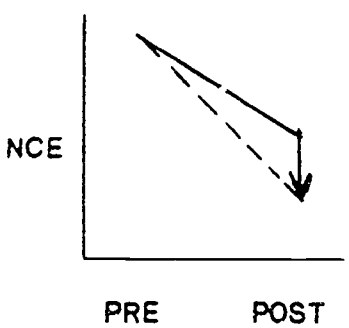
Model A:
not effective



Regular curriculum:
above ave. effect

Chapter 1:
not effective

Model A:
above ave. effect



Regular curriculum:
not effective

Chapter 1:
not effective

Model A:
very ineffective

Figure 2. Six possible Model A results, some of which are misleading.

Item #	In Regular Curriculum	In Chapter 1 Curriculum	Item Designation	Posttest Percent Correct	Inherent Difficulty
1	Y	Y	C	.40	.40
2	Y	Y	C	.78	.68
3	Y	Y	C	.80	.85
4	Y	N	R	.30	.40
5	Y	N	R	.68	.78
6	Y	N	R	.10	.20
7	N	N	OMIT	.20	.40
8	N	Y	OMIT	.50	.78
.					
.					
.					
20	Y	Y	C	.20	.15

Figure 3. Sample design.

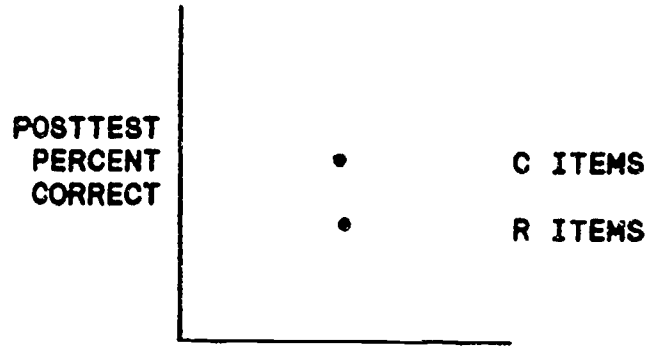


Figure 4. Schematic results from analysis I, two group means.

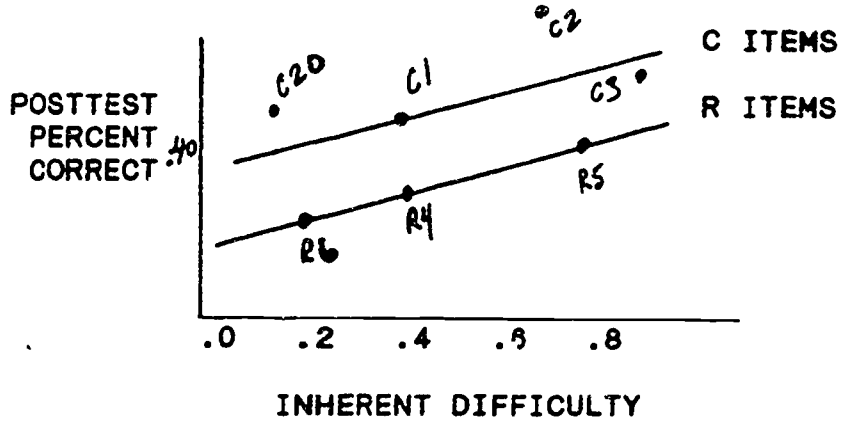


Figure 5. Schematic results from analysis II, inherent difficulty as covariate.

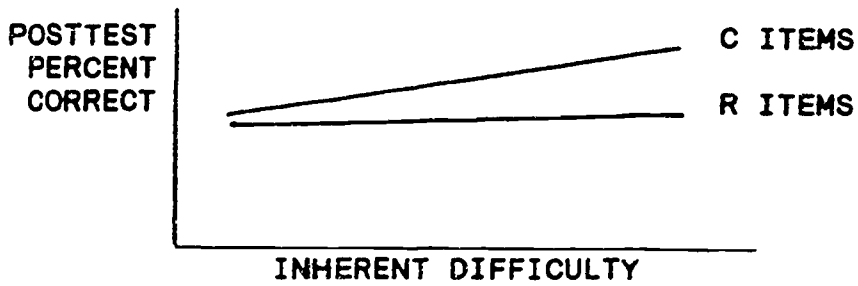


Figure 6. Schematic results from analysis III, linear interaction.

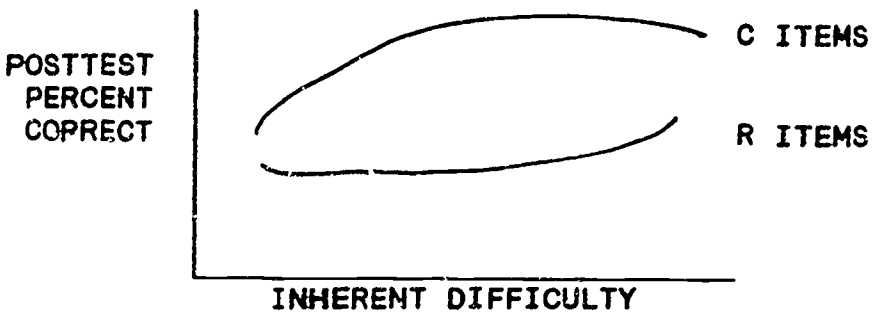


Figure 7. Schematic results from analysis IV, curvilinear interaction.

LEA	86-87	87-88	88-89	89-90
#1	5	5	5	5
#2	5	5	10	10
#3	2	2	2	2
#4	10	10	10	10
#5	5	5	1	1
#6	-2	-2	-2	-2

The Chapter 1 program in LEA #1 had a positive and consistent effect over the four years.
The Chapter 1 program in LEA #2 had a positive effect, moreso after the first two year.
The Chapter 1 program in LEA #3 had a consistent low positive effect in each of the four years.
The Chapter 1 program in LEA #4 had high positive effects in each of the last four years.
The Chapter 1 program in LEA #5 had a positive effect the first two years, but something happened in the last two years to eliminate the effect.
The Chapter 1 program in LEA #6 has had a consistent negative effect over the last four years.

Figure 8. Possible patterns of results from the one group posttest only design, along with possible interpretations.

Table 1. Percent Correct on STEELS Language Arts Items Included and Not Included in the A Priori Curriculum, 1987-88.

Grade	Items Included in A Priori		Items Not Included in A Priori		Probability of Difference
	Percent Correct	N	Percent Correct	N	
1	70.1	13	66.9	20	.009
2	73.3	18	71.9	23	.205
3	66.9	16	65.1	21	.124
All	70.1	47	68.2	64	.002

Note. Items were adjusted for overall difficulty.

Table 2. Percent Correct on STEELS Language Arts Items Included and Not Included in the A Priori Curriculum, 1988-89.

Grade	Items Included in A Priori		Items Not Included in A Priori		Probability of Difference
	Percent Correct	N	Percent Correct	N	
2	70.0	18	72.4	23	.72
3	70.8	16	64.5	21	.04
All	70.4	34	68.3	44	.12

Note. Items were adjusted for overall difficulty.

REFERENCES

Horst, D.P., Tallmadge, G.K., and Wood, C.T. A practical guide for measuring project impact on student achievement. Washington, D.C.: U. S. Government Printing Office, 1975 (Stock No. 107-080-01460).

McNeil, K., Berry, R., and Metze, B. (1988, July). Chapter 1 Basic Skills Final Evaluation Report, 1987-88 (REIS88-001-7). Dallas, Texas: Dallas Independent School District, Department of Research, Evaluation and Information Systems.

McNeil, K., Jones, K., Berry, R., Edoghotu, F., and Kane, R. (1989, July). Evaluation of the 1988-89 Chapter 1 Instructional Program (REIS89-001-5). Dallas, Texas: Dallas Independent School District, Department of Research, Evaluation and Information Systems.

McNeil, K.A., Kelly, F.J., and McNeil, J.T. Testing Research Hypotheses Using Multiple Linear Regression. Carbondale: Southern Illinois University Press, 1975.

Roberts, A.O.H. Out-of-level testing. In Evaluator's References: Title 1 Evaluation and Reporting System Vol 2. Washington, D.C.: U.S. Government Printing Office, 1981 (Stock No. 728-190-1792).

Ryan, J. Personal communication, 1980.