CONTROL TEST ITEMS:
A BASELINE MEASURE FOR EVALUATING
ACHIEVEMENT

Clinton B. Walker

CSE Report No. 108
March 1978

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

The work upon which this report is based was supported by Contract B6B-5526 issued by the Community Services Administration (CSA) to the Center for the Study of Evaluation. Points of view or opinions expressed do not necessarily reflect the policy of CSA and no official endorsement should be inferred.

CONTROL ITEMS: A BASELINE MEASURE FOR EVALUATING ACHIEVEMENT[1]

Clinton B. Walker

Student achievement is often the major concern in evaluations of educational programs. The evaluator gives a posttest of achievement to pupils in a given program, then tries to interpret the scores by comparing them with some type of performance baseline, norm, or expectancy. Two common types of baselines are provided by time series data and by independent control groups (Campbell & Stanley, 1963). In this report another type of comparison is discussed for which the name control test items, or simply control items, is coined. The following topics will be discussed in turn: rationale and precedents for the method, advantages and disadvantages, sources of possible control objectives, and unresolved issues.

Rationale and Background

The tension between advocates of norm-referenced tests (NRTs) and advocates of criterion-referenced tests (CRTs) provides a rationale for the control items technique. One objection to NRTs is that they measure skills or content which are outside of those actually taught in the classroom (GAO, 1977). This contention was dramatically documented for one evaluation by Tenenbaum and Miller (1977), who found that only one-third of the items in a particular NRT dealt with skills which were covered in the classes of a remedial program that was being assessed with that test. When an NRT includes a

[1] Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, March, 1978.

number of such "program irrelevant" items, an evaluator could do one of three things about them. First, (s)he could ignore the difference between the items that address program objectives and those that do not, and count all items together as evidence of a program's effectiveness. This approach to test content is the most common in current evaluations. Second, the evaluator could identify the two types of test items and score only those which the program addressed. Tenenbaum and Miller took this approach.

The prevalence of "program irrelevant" test items in NRTs is claimed to make such tests relatively insensitive to instruction. The merits of NRTs aside, this claim suggests a third approach to dealing with test questions which measure skills that are not intended outcomes of an educational program: to score them and use the data as a source of baseline or control information. For evaluating the effects of a curriculum, a test can be made of two types of items: ones that test explicit program objectives (program items) and ones that test explicit objectives that are not only beyond those of the program, but are also relatively insensitive to transfer from program instruction (control items). For example, fifth-grade students in a traditional math program could be tested on the program skills of basic operations, fractions, and decimals, and on the control skills of identifying the decimal equivalents of Roman numerals and base-four numerals. In a pretest/posttest design, the null hypothesis would be that there are no differences in growth on the two types of items. A finding of greater growth on the program items is a necessary but not sufficient condition for concluding that the program worked.

When we use control groups, we give one test of (presumably) program skills to two comparable sets of people, namely program and control students. In contrast, the method proposed here consists of giving one group of program pupils a test of two comparable sets of skills, namely program and control objectives. The familiar control group baseline is an estimate of how program pupils would have performed on program objectives without the benefit of the relevant instruction. The control items baseline is also an estimate of how program pupils would have performed on program objectives without the benefit of relevant instruction. In the traditional case one draws inferences from one group of pupils to another, but here the inference is from one group of skills to another.

Control test items are thus a conceptual analogue of control groups. Just as control and treatment groups need to be comparable, control and program test items need to be comparable in several respects. First, both sets of skills should be equally subject to spontaneous or artificial sources of seeming growth. If this condition is satisfied, then the posttest scores on control items give an estimate of how pupils would have performed on the program items if they had not received program instruction. Second, the two types of items should be equally learnable or sensitive to instruction. Item sensitivity is related to, but more than, mere item difficulty. It requires that floor and ceiling effects do not differ significantly for the two types of items and also that the control skills be as easy or hard to improve as the program ones. If this second type of comparability is present, posttest scores on the program items will give an estimate of how pupils would have performed on the control items if the control skills had been taught. An evaluator would thus look for an interaction--greater growth on program items than on control ones--rather than a main effect of item type.

Opinions differ on the need for a third type of comparability, which we call conceptual similarity. Objectives are conceptually similar if they are from the same content domain and reflect similar specific contents and processes. For a program objective on irregularly spelled words, a conceptually similar control objective could test irregularly spelled words that were not in the program vocabulary. This kind of similarity works against the need for control measures to be insulated from transfer from the program. Further study is needed to determine whether control skills from entirely different domains are theoretically meaningful.

Comparability of program and control items may be established by analysis of students' responses or by expert judgment. Data that have been used for levelling tests may confirm that program and control measures are comparable, but unless such data are carefully related to degree of instruction, they may be hard to interpret. Probably the best confirmation would come from a scaling study in which potential control and program skills were taught and tested. Expert judgments, either in the form of ratings or task analyses of the skills to be tested, could also be used to establish comparability.

The control test items method is in the tradition of research design as elaborated by Campbell and Stanley (1963). It combines basic concepts from that tradition, namely comparing treatment and control measures, with the concept of objectives-based testing, to provide a novel method of evaluating achievement. In the language of experimental psychology, the method involves presenting program pupils a "mixed stimulus list," where they respond to both treatment and control stimuli intermixed in one test form. The logic is that of construct validation, where treatment effects are the analogue of a construct.

The treatment is predicted to improve pupils' performance on treatment-focused measures and not on the treatment-irrelevant ones. In form the method is like that of divergent validation (Campbell & Fiske, 1959), but here it is treatments, not tests, that are to be validated by a divergence of scores on program and control measures.

Three more direct precedents for control test items have come to light since the present author began to develop the method. First, Millman and Gowin (1974) anticipated the method, suggesting that an educational treatment could be evaluated by comparing pupils' scores on treatment and non-treatment skills. Next, Lumsdaine (Hovland, Lumsdaine, & Sheffield, 1949) designed an achievement test with two types of items: items on the content of a training film and other items on the same general subject area that were not covered by the film. These other questions were used to establish the equivalence of the experimental and control groups so that differences in the groups' performance on the film-related questions could be ascribed to treatment effects. Finally, Chester Harris (personal communication) has proposed to measure the sensitivity to instruction of an individual test item by comparing changes in pupils' performance on the (instructionally-relevant) item with concurrent changes on a general I.Q. test item. In this case, the I.Q. question provides a baseline or control measure which is subject to the same artifactual sources of improvement as the target test item.

Potential Advantages and Disadvantages

When the two baseline devices of time series and control groups are already available, what could control test items add to the stock of evaluation tools? First, they give another indicator of program effects that can be used along with time series and control groups. When used in the same test with program items, control test items protect against the following artifacts: practice effects of repeated testing, maturation of the pupils, the general history of the program pupils (e.g., receipt of a school lunch program), and regression. Like control groups, control test items will give results that are susceptible to the artifacts of selection and mortality in the pupil population (Campbell & Stanley, 1963). Social action programs are complex enough to require a pattern of evidence to evaluate; control test items can be one of the multiple indicators in such a pattern, with or without control groups.

Second, control test items provide a method of comparison to use when time series cannot be used (e.g., when there are not enough data points in the series) and when control pupils are completely unavailable, available in insufficient numbers, or not clearly comparable to the treatment groups. In such cases it may be possible to estimate treatment effects solely by comparing pupils' performance on program and control items. To justify that conclusion, it is necessary to confirm that students' growth in program skills was not due to other sources of instruction on program objectives. This rival hypothesis of growth on program skills due to other sources may be weighed in the context of at least two types of other information: pupils' relative exposure to program and extra-program sources of instruction, and the past effectiveness (or in many cases ineffectiveness) of the other educational influences.

A third potential advantage of control test items is that they are not subject to some of the problems that control groups have. For example, treatments often leak into control sites, particularly when the control sites are within the same districts or even within the same schools as the program sites. Since control items give a within-test, within-pupils baseline, treatment and control measures of program pupils can be compared without regard to program leakage. Relative to control items, the analogue of program leakage into control sites is leakage of control objectives into program instruction. This reactive effect is a possibility, but it can be reduced or identified by various design and documentation devices.

A general objection to using control groups at all arises when there is not a specific control treatment: it is not clear what other treatment(s) the program treatment is being contrasted with. Students in the control sites are not put in suspended animation each day while program pupils are receiving the program; control pupils receive other experiences which vary from site to site. Experts disagree on whether such variability in control sites spoils the results from those sites. Since program implementation usually varies widely, this problem is not peculiar to control group designs. But given a well specified program to be evaluated, one advantage of the control items method is that it can provide a homogeneous, distinct control measure. That is, it can be a measure of specific behaviors that were clearly not taught in the program although they could have been.

Three limits or disadvantages of the control test item method are apparent. For one, they do not give the type of information that control groups do, namely, a comparison between existing programs. The population of potential control objectives might be conceived as an alternative, untaught program, but researchers who want a comparison of programs in place will not find that conception satisfying. If the test materials were accompanied by some type of comparative information, like norms, then the scores of program pupils could be related to that external standard. Some commercially available criterion-referenced testing systems do provide comparative data.

Next, since it is essential to have program and control skills clearly identified, and since it is essential that both sets of skills be sensitive to instruction, the control items method is more readily adaptable to evaluating objectives-based curricula and to doing so with item pools that were not developed for norm-referenced tests. This limitation is a relative one: it would be harder, but not impossible, to use the control items approach outside of an objectives-based teaching and testing context. Tenenbaum and Miller (1977) devised one method for doing so which involved relating individual NRT items to the content in program teachers' lesson plans.

Finally, the use of control items may reduce the reliability of the total scores on program items. The more control items there are in a test of fixed length, the fewer program items there can be. As the number of program items declines, so does the reliability of the program sensitive measure. However, this cost of using control test items may be negligible for tests that are constructed with item sampling techniques and multiple test forms.

Sources of Possible Control Objectives[2]

The art of selecting or discovering control objectives is only starting to develop, but there are some guidelines. First, look for skills which are relatively circumscribed or discrete. Such skills involve specific vocabulary or operations, such that they need to be studied directly in order to learn. This characteristic reduces the likelihood that the control measures will be subject to transfer from the program. Set theory notation may be an example. Next, look for skills which used to be taught and learned at the relevant level, but which are not now fashionable. Such skills would probably be comparable with program skills in the senses discussed above. Some objectives dealing with grammar fit this description. Third, look for skills which are traditionally taught at higher levels but which are not inherently harder to learn at the level tested. Identification of Roman numerals may be an example here. A fourth source is sometimes available: when a program objective samples only a part of a well-defined content area, like irregularly spelled words, control items may be selected from the excluded part of that content. A fifth possible source is available: objectives from outside the content domain of the program. For example, control items on identifying the muscles and bones of the body could conceivably be included in a test of program math objectives. Such objectives, however, may be too reactive and, as mentioned above, might have theoretical problems.

[2] The suggestions of CSE research staff members Meg Dotseth and Lee Trithart for this section are gratefully acknowledged.

Issues to be Resolved

Issues of sampling and formatting will not be discussed here except to raise some questions. In experimental research we randomly assign and sample people. In the case of control test items, the analogue would be random assignment and sampling of objectives in program and control conditions. But random assignment of objectives is not a sensible way to design curriculum, and random selection of program objectives is not the best way to design tests. Sampling issues include how to define the universe of control and program skills; how to sample objectives therefrom; how to judge the adequacy of sample size; and how or whether to generalize to the populations of objectives. After the population of possible control objectives has been narrowed by eliminating program objectives, objectives that are likely to exhibit positive transfer from the program, and objectives lacking comparability in the senses discussed above, that population may turn out to be quite small. Sampling issues interact with test formatting in that the number of items per objective on the test will be inversely related to the number of objectives sampled, assuming that test length is held constant. At one extreme would be a survey test covering many objectives with one item per objective. The formatting issue also relates to the choice of combining program and control items in one test as opposed to having two distinct full-length test forms, one of program items, the other of control items. A full-length "control test" might be a shock to a pupil, but it is a theoretical possibility which might serve as a heuristic for solving issues of sampling and formatting.

Afterword

Discussion of the control items method invariably elicits questions like these:

> What if some of the supposed control objectives get covered in the program?
>
> Isn't there likely to be a reactive effect on pupils and teachers of testing children on skills that are strange to them, like the control skills?
>
> What if the program and control items are not closely comparable on the various relevant dimensions?

Such questions can be especially valuable because they raise issues that are not unique to the control test item method. For each of the questions alone there is a parallel question about the traditional devices of norm-referenced tests and control groups, e.g.:

> What if the program leaks into the supposed control classrooms?
>
> Isn't there likely to be a reactive effect on pupils and teachers of testing children on skills that are strange to them, as is often the case with norm-referenced tests?
>
> What if program and control groups are not closely comparable on the various relevant dimensions?

Such questions bear on the most basic issues of how to get meaningful results and make meaningful comparisons at all. By pursuing such questions in this new context, the context of control test items, we may come to more powerful and practical solutions to some of the enduring measurement and design problems in evaluation.

REFERENCES

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand-McNally, 1963.

General Accounting Office (GAO). Report to the Congress: Problems and needed improvements in evaluating Office of Education programs. Washington, D. C.: U.S. Government Printing Office, 1977. Document No. HRD-76-165.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. Experiments on mass communication. Princeton, N. J.: Princeton U. Press, 1949.

Millman, J., & Gowin, B. Appraising educational research: A case study approach. Englewood Cliffs, New Jersey: Prentice Hall, 1974, p. 46.

Tenenbaum, A. B., & Miller, C. A. The use of congruence between test items in a norm-referenced test and the content in compensatory education curricula in the evaluation of achievement gains. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April, 1977.