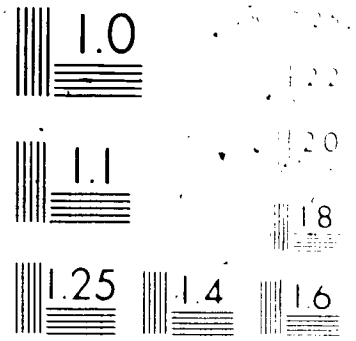
1.0
1.1
1.25
1.4
1.6
1.2
2.0
1.8

DOCUMENT RESUME

ED 169 127 TM 008 649

AUTHOR Horst, Donald P.; Fagan, Barbara M.
TITLE Common Evaluation Hazards. ESEA Title I Evaluation and Reporting System. Technical Paper No. 14.
INSTITUTION RMC Research Corp., Mountain View, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE Oct 76
NOTE 16p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Gains; Achievement Tests; Elementary Secondary Education; *Evaluation Methods; Evaluation Needs; Models; Post Testing; Pretesting; *Research Design; *Research Problems; Scores; Scoring; *Student Testing; *Testing Problems; Test Interpretation; Test Selection
IDENTIFIERS Elementary Secondary Education Act Title I

ABSTRACT

Twelve common errors which can invalidate an otherwise sound evaluation are identified, and ways to avoid them are presented. The hazards are: (1) grade-equivalent scores; (2) inappropriate statistical adjustments with nonequivalent control groups; (3) administering norm-referenced tests at inappropriate times of the school year; (4) inappropriate grade levels of tests; (5) missing scores; (6) noncomparable treatment and control groups; (7) use of a single set of scores for both selecting and pretesting participants; (8) constructing a matched control group after the treatment group has been selected; (9) careless administration and scoring; (10) changing instruments between pretesting and post testing; (11) inappropriate formulas such as IQ or pretest scores to generate no-treatment expectations; and (12) mistaken attribution of observed gains. A table indicates which hazards present the biggest threat to the validity of three Title I evaluation models: control group, norm group, and special regression. (CP)

COMMON EVALUATION HAZARDS

ESEA Title I
Evaluation and Reporting System

Technical Paper No. 14

Donald P. Horst
Barbara M. Fagan

October 1976

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

RMC Research Corporation
Mountain View, California

When used correctly, tests provide the user with sound, interpretable results. When tests are used inappropriately in evaluations, results are often inconclusive. In a review of 2,000 evaluation reports from educational projects, Horst, Tallmadge, and Wood (1975) identified common errors or hazards in evaluations and in the use of tests. The hazards and ways to avoid them are presented below. Any one of them can invalidate an otherwise sound evaluation, and should be avoided.

Hazard 1: The Use of Grade-Equivalent Scores

Grade-equivalent scores provide an insensitive, and, in some instances, a systematically distorted, assessment of cognitive growth. The concept of a grade-equivalent score is misleading--for example, a grade-equivalent score of five attained by a third grader on a math test does not mean that he knows fifth-grade math. Possibly he can do third-grade math as well as the average fifth grader, but it is likely that no fifth-grade students have ever taken the third-grade level of the test.

The use of grade equivalents for evaluation purposes creates a second problem in that they do not form an equal-interval scale and should never be averaged. Finally, grade equivalents are constructed based on the assumption that growth occurs at the same rate throughout the school year. Research has shown, however, that learning typically does not follow this regular pattern and, whenever this is the case, gains measured in grade equivalents will be artificially inflated or reduced. For a complete discussion of their problems, see Technical Paper No. 1 entitled What's Bad about Grade-Equivalent Scores.

Hazard 2: The Use of Inappropriate Statistical Adjustments with Nonequivalent Control Groups

There are several statistical procedures that are widely used in an attempt to compensate for initial differences between treatment and control groups. Some are legitimate while others are not. Making between-group comparisons using either "raw" gain scores or "residual" gain scores falls into the latter category. Both procedures should be scrupulously avoided.

A raw gain score is simply the difference between a pre- and a posttest score and reflects the gain made between testings. It is argued that, although two groups may have been somewhat different in terms of initial achievement levels, their expected gains would be roughly comparable after the same educational treatment. This would be true, however, only when each group's posttest standard deviation is the same as its pretest standard deviation. Where the posttest standard deviations are larger than those of pretest scores, a raw gain score analysis will systematically underestimate treatment effects. Conversely, the procedure will systematically overestimate treatment effects where the standard deviations of pretest scores exceed those of posttest scores.

A residual gain score is the difference between an actual posttest score and a posttest score estimate derived from the combined treatment and control group regression line. Presumably the mean residual gain score for a group which received an effective treatment would be positive while that for the control group would be negative. Also, the sum of the absolute values of the two differences would provide an index of the size of the treatment effect. Unfortunately, it can be shown algebraically that a residual gain

score analysis always underestimates the size of the treatment effect except where the groups' pretest scores are equal. Furthermore, the amount of understatement is directly proportional to the size of the initial difference between groups.

There are other factors, such as how the treatment and control groups were formed, which determine the appropriate adjustment procedure to compensate for their initial differences. Refer to Technical Paper No. 12 entitled Statistical Adjustments for Nonequivalent Control Groups for a more complete discussion of this topic. Here it is sufficient to point out that neither raw nor residual gain score adjustments is adequate.

Hazard 3: The Use of Norm-Group Comparisons with Inappropriate Test Dates

In norm-referenced evaluations, tests should be administered at nearly the same time as the test publisher tested the norm group. When control groups are available, few evaluators would consider testing the treatment and control groups more than a few days apart. When norms are used as a substitute control group, this same consideration needs to be given to test dates.

Treatment group students should be tested within two weeks of the midpoint of the interval during which the normative data were collected. Testing within six weeks of empirical normative data points is permissible if linear interpolations or extrapolations of the normative data are made. Tests that provide normative data for only one point in the year should not be used in fall-to-spring norm-referenced evaluations.

Hazard 4: The Use of Inappropriate Levels of Tests

If most of the pupils achieve very high or very low test scores, the level of the test may

be inappropriate for assessing their performance. If pupils encounter the test floor at pretest time or the ceiling at posttest time, treatment effects will be underestimated. Conversely, if the ceiling is encountered on the pretest or the floor on the posttest, gains will be overestimated. Ideally, students should score in the middle of the range of possible raw scores.

Test levels should be selected on the basis of the achievement levels of the students, not on the basis of their grade in school. In most cases, the nominally recommended test level of one level below will be suitable for testing Title I students. See Technical Paper No. 6 entitled Out-of-Level Testing for additional information on this topic.

Using a test level other than that nominally recommended for a particular grade is likely to mean that norms tables for the tested students are not included in the test manual. However, it is not meaningful to assess either status or gains by comparisons with students at a different grade level. The status of a sixth grader should be assessed using sixth-grade norms even if he is tested with a fourth-grade test. Most major test publishers, fortunately, have interlocked their test levels by providing an expanded standard score scale which enables the determination of score equivalencies between adjacent test levels. These scores make it possible to predict from a pupil's score on one test level how he would have scored on the next higher or lower level, thus providing access to the in-level norms.

Hazard 5: Missing Test Scores

Analyses of evaluation data should be based only on those students with both pre- and post-

test scores. Interpretation of these data, however, should take into account the characteristics of the students who dropped out, entered late, or graduated from the project. For example, if all of the lowest scoring students on the pretest dropped out before posttest time, the average posttest score would increase with respect to the pretest scores simply because of the missing students. This increase could be misinterpreted as a gain. Likewise, if the high-scoring students graduated from the group, the mean posttest score would be artificially deflated.

To avoid this hazard, every effort must be made to obtain pre- and posttest scores for each project participant, and to base comparisons on those students for whom both scores are available. Data from students having only pretest or only posttest scores must be carefully examined to see if they differ in some systematic way from the data of students having both pre- and posttest scores. A description of any of these differences should be included in the evaluation report.

Hazard 6: The Use of Noncomparable Treatment and Control Groups

This hazard is closely related to Hazards 2 and 8. In conventional experimental designs, treatment and control groups should be similar in all educationally relevant respects before the treatment begins. Groups which differ in terms of pretest scores present an obvious source of bias. Other more subtle factors such as differences in age, sex, race, or socioeconomic status can also exert strong biasing influences and should be avoided. Nonvolunteers should never be used as controls for pupils who volunteered (or were volunteered by their parents) for a particular instructional treatment.

Whenever possible, students should be assigned to treatment and control groups on a random basis. For example, with a semester-long reading program, pupils could be randomly assigned to first- or second-semester groups. For the first half of the year, one group would serve as the control group for the other, but both groups would ultimately receive equal amounts of the treatment.

In some cases, pre-existing groups will be enough alike so that they can appropriately be considered equivalent to random samples from a single population. In other cases, a control group will be known to differ systematically from the treatment group. Where the difference is small, the control group model may still provide the best method of evaluating the project, and statistical adjustments can be made to compensate for between-group differences (see Technical Paper No. 12, entitled Statistical Adjustments for Nonequivalent Control Groups). Where the differences are large, however, there is no way in which a noncomparable control group can provide an accurate estimate of how well the treatment group would have done without the treatment.

Hazard 7: The Use of a Single Set of Test Scores for Both Selecting and Pretesting Participants

When students are selected for participation in a special group because they obtained relatively high or relatively low scores on some test, use of these scores as pretest measures invalidates any kind of norm-referenced evaluation. This problem stems from what is known as "statistical regression," "regression toward the mean," or simply, the regression effect. For a discussion of this topic, refer to Technical Paper No. 3 entitled The Regression Effect.

If low-scoring students are retested on the same or a comparable test, they will score higher on the average, while an initially high-scoring group will score lower. The result is that low-scoring groups appear to learn more from a special program than they actually do, while gains in special programs for high-scoring students may be obscured.

To avoid this hazard, students should be selected for participation in a special treatment based on one set of test scores and then be pretested using an alternate form of the same test or a different test. A perfectly legitimate alternative is to base student selection on teacher recommendations or classroom grades.

Hazard 8: Constructing a Matched Control Group After the Treatment Group Has Been Selected

Finding "matches" for treatment participants in some other group is a fundamentally unsound practice. Unless they and the treatment pupils are equally representative of the groups from which they are drawn, statistical regression will act differentially on the two groups and artificially inflate the apparent gains of one group with respect to the other.

In the most common situation, the group(s) from which the matching control pupils are drawn will be higher achieving than those from which the treatment group pupils are selected. Consequently, the control group pupils will be farther below the mean of the group(s) to which they belong than the treatment group children. On retesting they will thus show greater statistical regression and their posttest scores will be too high to serve as a no-treatment expectation for the Title I participants.

The correct procedure for establishing matched control groups is to do the matching first and then assign members of each pair randomly to the treatment or the control group. That is, a large group of students, all eligible to be in the project, must be available. The first step is to divide the group into matched pairs based on test scores, ethnic background, sex, etc., so that the two members of each pair are as similar as possible. Then, after the matching process is complete, some random procedure such as flipping a coin should be used to decide which member of each pair goes into the treatment and which into the control group.

Hazard 9: The Careless Administration and Scoring of Tests

Testing must be accomplished with scrupulous attention to detail. For most evaluation models, the primary requirement is that treatment and control or comparison groups be tested in exactly the same way. Minor variations from the procedures described by the test publisher are permissible. In norm-referenced evaluations, treatment groups should be tested in the same way as the students in the norm group. This requirement means that procedures outlined by the test publisher must be followed precisely.

Problems arise if tests are administered or scored in an inconsistent and careless manner. If there are differences in the ways in which the test takers and the norm group students are tested or if there are differences in the procedures, conditions, and scoring at pretest and posttest times, then it is impossible for the resulting data to be accurately interpreted. There are no statistical manipulations that can compensate for mistakes made in administering or scoring a test.

To avoid this hazard, the following steps should be taken:

1. Test procedures must be orderly and accurate if scores are to be meaningful.
2. Test administration and scoring procedures must be exactly the same for the treatment group as for the control, comparison, or norm group used to generate the no-treatment expectation. Testing treatment group pupils in exactly the same way as pupils in the norming sample means following the test publisher's directions in every detail.
3. The procedures, conditions, and scoring methods used during posttesting must be exactly the same as those used during pretesting.

Hazard 10: The Use of Different Instruments for Pretesting and Posttesting

In the norm-referenced design, it is not advisable to change tests between pre- and posttesting because there is no adequate way to compare pretest scores on one test with posttest scores on a completely different test. Since each test publisher follows slightly different norming practices, it is likely that one test's norms will be slightly "easier" than another's. This difference does not matter if the same test is used both pre and post but could magnify or obscure real gains if changes were made. While it is not essential to use the same form and level of an achievement test pre and post, this practice is also recommended.

Some tests have been developed so that the lower levels are intended for use at the end of one grade and the beginning of the next. In these instances, to use the same form and level of test for fall pretesting and spring posttesting, it

will be necessary either to pretest or posttest out-of-level. In some grades where spring-to-spring or fall-to-fall evaluations are conducted, it may be necessary to change test levels in order to avoid ceiling or floor effects; unfortunately, this practice will introduce an unknown amount of error into the measure of gain.

Hazard 11: The Use of Inappropriate Formulas to Generate No-Treatment Expectations

Many projects use an unrealistic theoretical model or formula to calculate "expected" posttest scores from IQ or other pretest scores. If students do better than the calculated expectation, the project is considered a success.

Many methods have been devised for calculating performance-level expectations which rest on untenable assumptions. Neither IQ scores nor grade-equivalent scores should be used to generate no-treatment expectations. For example, a student who has gained .7 years per year, on the average, since beginning school, is presumed to continue at the same rate unless a special program increases his rate. Unfortunately, grade-equivalent gains measured from fall to spring will usually exceed this rate--even for typical Title I children--and treatments will appear to be more effective than they really are.

In norm-referenced models, no-treatment expectations should be generated solely from empirical percentile norms tables. When control groups are used, the actual posttest scores of these groups provide the proper basis for evaluating treatment effects. In the special regression model, a regression line based on comparison group data can be used to estimate the posttest scores.

Hazard 12: Mistaken Attribution of Causality

Observed gains may have resulted from the Title I treatment, but there are always plausible alternative explanations. The plausibility of these alternative explanations should be carefully examined before evaluation results are attributed to project impact, as evaluation hazards are often the cause of apparent gains or non-gains.

Sometimes project participants learn substantially more than would have been expected, but the project, per se, is not responsible. Instead, the gains could be a result of the Hawthorne effect (Whitehead, 1938) in which special project participants do well simply because they are getting special attention. The nature of the treatment may not necessarily be important. An opposite result may follow from a John Henry effect (Saretsky, 1972). In this case, comparison group students work extra hard to prove that they are just as good as project students.

Other likely causes of misleading gains are unrecognized "treatments" which have nothing to do with the project. Most school systems are in a constant state of flux with multiple changes every year. Changes in school programs, personnel, facilities, class sizes, community characteristics--any or all of these factors can affect student performance. Also, the true source of achievement gains is sometimes improperly identified because children are involved in more than one treatment. Under these conditions, it is impossible to determine causality in an unambiguous manner.

* * *

The table below indicates which hazards present the biggest threat to validity of the Title I evaluation models.

EVALUATION HAZARDS BY MODELS

	Control Group Model	Norm Group Model	Special Regression Model
1. Grade-equivalent scores	X	XX	X
2. Inappropriate adjustments	X		
3. Norm-referenced testing on inappropriate dates		X	
4. Inappropriate test levels	X	X	X
5. Missing test scores	X	X	X
6. Noncomparable groups	X		
7. Selection based on pretest scores		X	
8. Post-hoc matching of groups	X		
9. Careless testing	X	X	X
10. Noncomparable pre- and posttests		X	
11. Inappropriate posttest estimates	N/A	N/A	N/A
12. Mistaken attribution of causality	X	X	X

REFERENCES

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide for measuring project impact on student achievement. Washington, D.C.: U.S. Government Printing Office. (Stock No. 017-080-01460).

Saretsky, G. The OEO P.C. experiment and the John Henry effect. Phi Delta Kappan, 1972, 579-581.

Whitehead, T. N. The industrial worker. Vol. 1. Cambridge: Harvard University Press, 1938.

ADDITIONAL READING

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide for measuring project impact on student achievement. Washington, D.C.: U.S. Government Printing Office. (Stock No. 017-080-01460).

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1967.

Tallmadge, G. K., and Horst, D. P. A procedural guide for validating achievement gains in educational projects. Washington, D.C.: U.S. Government Printing Office. (Stock No. 017-080-01516).