DOCUMENT RESUME

ED 126 135 95 TM 005 391

AUTHOR Horst, Donald P.; Tallmadge, G. Kasten
TITLE A Procedural Guide For Validating Achievement Gains in Educational Projects. Monograph Series on Evaluation in Education, No. 2.
INSTITUTION RMC Research Corp., Los Altos, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE 76
CONTRACT OEC-0-73-6662
NOTE 103p.; For a related document, see ED 096 344
AVAILABLE FROM Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 ($2.10)

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; Compensatory Education Programs; Control Groups; Criteria; Criterion Referenced Tests; Data Collection; Decision Making; *Demonstration Projects; Educational Programs; Grade Equivalent Scores; *Guides; Mathematical Models; Measurement Techniques; Models; Norm Referenced Tests; Norms; *Program Effectiveness; *Program Evaluation; Research Design; Research Problems; Selection; Standardized Tests; Statistical Analysis; Tests of Significance; *Validity
IDENTIFIERS Percentile Norms

ABSTRACT

The orientation of this report is that of identifying educational projects which can be considered clearly exemplary. The largest section consists of a 22-step procedure for validating the effectiveness of educational projects using existing evaluation data. It is not intended as a guide for conducting evaluations but rather for interpreting data assembled by others using a wide variety of experimental and quasi-experimental designs. As such, its coverage is not restricted to "good" designs. It encompasses all of the commonly employed evaluation models, but is not so much concerned with assessing the relative usefulness of various designs as with the deficiencies and hazards inherent in each of them. It also offers suggestions for correcting those results when certain measurement or analysis principles have been violated. Included as appendices are a discussion of the issues surrounding use of criterion-versus norm-referenced tests, description of the logic and mathematical structures of certain regression models, and an overview of the hazards associated with the use of percentiles and grade equivalent scores to describe academic performance. (Author/DEP)

A PROCEDURAL GUIDE FOR VALIDATING ACHIEVEMENT GAINS IN EDUCATIONAL PROJECTS

Number 2
in a Series of
Monographs on
Evaluation in
Education

A PROCEDURAL GUIDE FOR VALIDATING ACHIEVEMENT GAINS IN EDUCATIONAL PROJECTS

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
David Mathews, *Secretary*
Virginia Y. Trotter, *Assistant Secretary for Education*
Office of Education
T.H. Bell, *Commissioner*

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.

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1976

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price $2.10

FOREWORD

This is the second in the Office of Education's series of evaluation handbooks. It complements the first* by approaching the problem from a different viewpoint--that of the interested party reviewing evaluation results and selecting exemplary projects based on them. Written by G. Kasten Tallmadge and Donald P. Horst of the Mountain View, California Office of RMC Research Corporation, it is a product of contract OEC-0-73-6662 entitled, "Planning Study for the Development of Project Information Packages for Effective Approaches to Compensatory Education."

Review and appraisal of an evaluation's procedures are presented in a series of steps. The handbook thus leads the reader systematically to a judgment of whether or not the evaluation's results are valid. It also offers suggestions for correcting those results when certain measurement or analysis principles have been violated. Included as appendices are sample project summary worksheets, a discussion of the issues surrounding use of criterion-versus norm-referenced tests, description of the logic and mathematical structures of certain regression models, and an overview of the hazards associated with the use of percentiles and grade equivalent scores to describe children's academic performance.

Other handbooks forthcoming in the series organized by the Office of Planning, Budgeting, and Evaluation will discuss procedures for using criterion-referenced tests in evaluation, for assessing children's affective growth, for estimating standard replicable project costs, for evaluating non-instructional project components, etc.

Janice K. Anderson
Office of Planning, Budgeting,
and Evaluation
U.S. Office of Education

* A Practical Guide to Measuring Project Impact on Student Achievement, Donald P. Horst, G. Kasten Tallmadge, and Christine T. Wood, RMC Research Corporation, Mountain View, California, 1975. Government Printing Office Stock Number 017-080-01460, $1.90.

ACKNOWLEDGMENTS

The present version of this document is the second major revision of a report first published in October, 1973. Both the original report and the first revision owed much to Edward B. Glassman of the U. S. Office of Education's Office of Planning, Budgeting, and Evaluation (OPBE). As Project Officer for the contract under which the report was developed, he deserves credit for originally recognizing the need for such a guidebook. The authors are also indebted to him for his many thoughtful suggestions and for those he solicited from his professional colleagues throughout the original writing and revision processes.

The present revision owes its origin to Janice K. Anderson, also of OPBE, who has taken on responsibility for the entire series of Monographs on Evaluation in Education. We are indebted to her for her many good ideas and for the encouragement she provided in getting us to think through the issues one more time.

We are also indebted to Paul Horst for the very great assistance he provided with Appendix C of the report and for his frequent additional comments and suggestions.

Many other members of the RMC Research Corporation staff also helped in various capacities, and we are most grateful to all of them.

G. Kasten Tallmadge
Donald P. Horst

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

I. INTRODUCTION

This report was developed in conjunction with Contract No. OEC-0-73-6662 entitled, "The Development of Project Information Packages for Effective Approaches in Compensatory Education." As its name implies, the contract effort was primarily focused on packaging concepts and procedures which would facilitate the replication of sound educational practices. There was great concern, however, that the projects selected for replication should indeed be exemplary in producing significant cognitive achievement benefits.

Because the selection process was to be based on existing data derived from a wide variety of experimental and quasi-experimental evaluation designs, it was necessary not only to establish criteria for the statistical and educational significance of achievement gains but also to define procedures for verifying that these criteria were met. This latter task was not regarded lightly, but it was, the authors felt, something which could be accomplished in a straightforward manner by borrowing liberally from the work of Campbell and Stanley (1963) and others. It did not seem likely that much original work would be required, or that this report would contain any significant information not already present in widely read evaluation texts. These initial impressions, however, were quickly to be rejected.

It was not long after work on the validation procedure began that it became necessary to put aside the well documented issues of experimental design and statistical inference and to probe the netherworld intricacies of achievement test scores and normative data. Facts which appeared to undermine the validity of inferences drawn from nearly all locally conducted evaluations quickly came to light as this exploration proceeded. The problems were so fundamental that the authors found it hard to believe they were not well known--yet

they were able to find little in the literature which was more than marginally relevant.

Before they started work on the validation procedure, the authors considered themselves reasonably sophisticated in both the theory and practice of educational evaluation. There were, however, a number of details which had escaped their attention. They were not aware, for example, that a child scoring in the lowest quartile of the national distribution could make gains greater than month-for-month over an entire school year and end up farther below the norm than he began. They did not know that a fiftieth-percentile third grader could be 2.5 months below grade level in reading--or that an educational program could appear highly successful if the pre- to posttest interval spanned the twelve months from 1 May to 1 May but would resemble an instructional disaster if pupils obtained the same scores on tests administered one day earlier.

These outrageous incoherencies were just a few of the "horror stories" uncovered in the course of routinely examining real-world evaluation studies. The sad part was that these or similar irrationalities were so pervasive that not a single evaluation report was found which could be accepted at face value! Even more disheartening--many of these evaluations followed procedures officially sanctioned by one or more presumably authoritative groups of experts.

With each new discovery it became increasingly clear that this report would have new things to say and would have significant implications beyond the scope of the effort which spawned it. For this reason, it has undergone several revisions intended to increase its general usefulness. One significant change involved removing as much as possible of the material which dealt with project selection criteria unrelated to cognitive achievement benefits. Discussion of these criteria (cost, availability, and replicability) was clearly specific to the contract effort and appeared to detract from the usefulness of the report for a broader audience.

While the coverage of the report has changed somewhat from earlier

versions, its format remains the same. The largest section of the report consists of a 22-step procedure for validating the effectiveness of educational projects using existing evaluation data. It is not intended as a guide for conducting evaluations but rather for interpreting data assembled by others using a wide variety of experimental and quasi-experimental designs. As such, its coverage is not restricted to "good" designs. It encompasses all of the commonly employed evaluation models, but is not so much concerned with assessing the relative usefulness of various designs as with the deficiencies and hazards inherent in each of them.

One additional point should be mentioned here. The orientation of this report is that of identifying educational projects which can be considered clearly exemplary. Unfortunately, in minimizing the probability of classifying an unsuccessful project as successful, the decision-tree procedures somewhat increase the probability of rejecting projects which may really be successful. If the goal were to identify unsuccessful projects for the purpose of terminating them rather than successful projects for replication purposes, a different orientation would be more appropriate.

II. PRELIMINARY SCREENING OF CANDIDATE PROJECTS

The process of selecting and validating exemplary educational projects is viewed as iterative in nature with each criterion area examined at several preliminary levels before analysis is undertaken at the depth which will ultimately be required. The specific steps to be taken and the criteria to be used will vary as a function of each study's particular objectives. The variations, however, should not represent major departures from the general strategy which was employed in selecting exemplary compensatory education projects for packaging. This strategy is described below.

The process began with defining the population from which projects were to be drawn, assembling a list of candidate projects, and soliciting available documentation from each of them. When these tasks were completed, the investigators had in their possession an incomplete collection of reports, data, and promotional literature on each candidate project.

Winnowing this information, identifying and obtaining needed supplementary data, and weighing the resulting evidence was a complex task. It required a substantial investment of effort including mail and telephone communication with project personnel and usually at least one site visit. Typically, it was not feasible to apply the entire process to all candidate projects, and some preliminary screening procedures were required. Projects which passed the preliminary screening criteria were considered "possible" candidates for validation, and all criterion areas were then systematically investigated in greater depth. When there was doubt as to whether or not a project had met one of the preliminary criteria, the project was not rejected immediately. Rather, attention was focused on the specific criterion in question so that definitely unsuitable projects could be identified and rejected with a minimum of superfluous effort.

Appendix A contains a set of worksheets which were developed to facilitate the preliminary screening of compensatory education projects which were candidates for exemplary status. While the specific criteria applied to this screening effort may not be widely applicable without modification, the worksheets should serve as useful models for any similar types of screening.

The first worksheet was completed for every candidate project and provided a record of the disposition of the project. The first two sections, "Description" and "Prerequisites," were completed as the first step in processing information received from a project. Information under these headings served to verify that the candidate project did indeed come from the population being considered. The third heading, "Final Assessment" was used later to summarize the results of the investigations in each of the four major criterion areas.

The second worksheet, "Preliminary Screening Criteria" comprises a checklist which was used for all projects which met the prerequisites. Checks were made whenever it was possible to determine that a criterion had been met. Conversely, if it could be determined that one of the criteria was not met, the project was immediately rejected and no effort was spent examining other areas. Where doubt existed, efforts were focused on the questionable areas one at a time until either it was determined that all criteria were met or the project was rejected.

The third worksheet, entitled "Analysis of Project Evaluation," was used to describe the tryout design in such a way as to summarize the evidence of effectiveness and provide a context for its interpretation.

The use of forms such as those included in Appendix A for summarizing and recording preliminary screening information may give the misleading impression that the screening process is quite rigorous. In fact, it is no more than a coarse grouping procedure whereby educational projects are categorized as (a) apparently meeting the selection criteria, (b) apparently not meeting the selection criteria, or (c) can't tell. Even the distinction among these groups is not at

all clear-cut in the effectiveness area where misuse of experimental designs and statistical procedures is quite common and affects results in ways that are not easily decipherable.

It was decided that the detailed validation procedures would be applied solely to projects which appeared, on the basis of preliminary screenings, to meet the selection criteria. Only if the number of such projects which survived validation was inadequate would it be necessary to dip into the "can't tell" category. At that point, validation procedures would be applied to those projects which the investigators felt were most promising based on whatever circumstantial evidence they could assemble.

This process would continue, one project at a time, until either the "quota" was filled or until it became clear that the original classification had been excessively optimistic and that the probability of finding additional successes was so remote as to suggest abandoning the search.

III. EVALUATING PROJECT EFFECTIVENESS

Assessing the effectiveness of an educational project presents an intrinsically difficult problem. The evaluator faces many pitfalls which may be broadly categorized as relating to measurement, experimental design, or statistics. Hazards exist in each of these areas which may completely invalidate any inferences which might be drawn from the data regarding project impact.

Conventions for experimental design and associated statistics have been developed to deal effectively with evaluation problems in controlled experimental settings. Standard reference books describing these conventions are widely available (e.g., Winer, 1971) and are well known to most evaluation specialists. Unfortunately, in the real world of education it is often impossible to employ rigorous techniques, and it is extremely rare to find a compensatory education project which satisfies all, or even most of the fundamental principles of good research design. The problem is so widespread, in fact, that if one were to reject all projects with less-than-ideal evaluations, the possibility of finding even a few exemplary projects would be extremely remote.

Many of the weaker designs have been discussed at length by Campbell and Stanley (1963) along with the "threats to internal and external validity" associated with each. These authors, however, have hardly touched upon the equally important and related problems of educational measurement. Scoring, scaling, and norming considerations are fundamental to all educational evaluations and are particularly critical to those designs which employ non-equivalent comparison groups or no comparison group at all.

The extent and complexity of the experimental and measurement problems inherent in evaluation call for a systematic procedure for reviewing project evaluations, for identifying and assessing the impact of their shortcomings, and for making reasonable judgments regarding

project effectiveness while carefully weighing all relevant factors. To meet this need, a 22-step decision tree was developed. The decision tree was designed to insure appropriate consideration, in any evaluation study, of each of the threats to valid inference discussed by Campbell and Stanley (1963) relevant to the specific design employed. It also focuses attention on problems related to whether comparisons are made against control groups or are norm-referenced, the type of scores on which statistical operations are performed (raw, standard, scale, percentile, grade-equivalent), and the bases on which treatment-control (or norm group) comparisons are made (posttest scores, adjusted posttest scores, gain scores, etc.).

A procedure of this type cannot, of course, be applied in a vacuum. It must be tied to pre-established criteria to which each judgment can be related. Principal among these criteria are (a) the minimum increment of cognitive benefit which will be considered educationally significant and (b) the minimum non-chance probability level which will be accepted as statistically significant.

It should be pointed out that the establishment of criteria for educational and even statistical significance is a matter of policy decision-making and has only tenuous ties to "science." While it is clear, for example, that the goal of compensatory education is to raise the achievement levels of disadvantaged children from some starting point to an end point which is closer to the national norm, the amount of gain required for projects to be considered exemplary is almost entirely a matter of opinion. The only scientific issue is that of selecting or developing a suitable metric for quantifying the cognitive gain criterion.

The use of grade-equivalent scores has appeared to offer a convenient solution to the problem. It is intuitively logical that, regardless of how far below the national norm a child may be, if he makes gains which are greater than month-for-month he will improve his status. It is also intuitively logical that if he makes gains which are less than month-for-month, he will fall farther behind the national norm. Thus it has been common practice to assess cognitive growth in terms of

grade-equivalent gains per month of project exposure and to accept gains equal to or greater than month-for-month as educationally significant. Unfortunately, this convention is fundamentally unsound and often leads to incorrect inferences about the impact of special instructional projects.

Because cognitive growth is not a linear function of time either between or within years, because test publishers do not collect enough normative data to construct more meaningful raw-to-grade-equivalent-score conversion tables, and because a lot of interpolation, extrapolation, and curve-smoothing is always involved, grade-equivalent scores simply do not behave in a fashion which is consistent with intuitive or logical expectations. These and other technical problems associated with grade-equivalent scores and grade-equivalent gains are discussed in detail later in this report, and examples of some of the incoherencies which actually occur in real-world situations were presented in the Introduction. Here it is sufficient simply to say that such scores do not provide a suitable medium for measuring the achievement gains that may result from compensatory education projects.

Even if grade-equivalent scores possessed the characteristics which they are typically presumed to have, the month-for-month measure of effectiveness would be deficient in that it would systematically discriminate against projects serving those most in need of compensatory programs. This systematic bias stems from the all-but-trivial fact that increasing an achievement growth rate from 0.9 to 1.0 month-per-month requires less remediation than raising one from 0.7 to 1.0. A more equitable measure would be one which is independent of the initial degree of disadvantagement of the children being served.

In order to be independent of initial achievement status, any measure of gain must be expressed in terms of an equal-interval scale, i.e., the units of the scale must be the same size over the entire range of scale values so that a gain of five points represents exactly the same amount of cognitive growth regardless of whether it occurs at the low end of the scale, the middle of the scale, or the high end.

Normalized standard scores comprise such a scale and thus provide an equitable metric for quantifying gains. There is another problem in quantifying gains, however, which relates to the non-comparability of information derived from one scale of normalized standard scores with that derived from another.

A standard score is simply the difference between a particular "observed" score and the mean score of the total group tested, expressed in standard deviation units.

$$\text{standard score} = \frac{X_i - \bar{X}}{s_x}$$

As such, its value (size) depends on both the mean and the standard deviation of the particular group which was tested. Different groups of course, can be expected to have different mean scores and different standard deviations; thus there will be no comparability between standard scores or standard score gains from group to group.

To solve the comparability problem, it is only necessary to use standard scores which are referenced to the mean and standard deviation of a nationally representative sample rather than the values derived from the particular group tested. If, for example, several different groups of children were tested at the beginning of third grade, the scores of each child could be expressed as deviations from the national average for beginning third graders divided by the standard deviation of the national population of these children. Scores derived in this way would provide a suitable metric for quantifying gains and would also enable equitable comparisons of gains to be made among projects serving children with different degrees of initial disadvantagement.

These considerations led the authors to advocate the use of standard score gains referenced to the national norm as the medium in which to cast whatever definition of educational significance might be decided upon. Subsequently, a gain of one-third standard deviation with respect to the national norm was chosen (on somewhat arbitrary grounds) as the criterion to be used in the national packaging effort for determining exemplary status. In that study, for a project to be

considered for packaging, the mean posttest standard score of project participants had to be one-third standard deviation higher with respect to the national norm than the mean pretest score of the same children.

Criteria for educationally significant gains will vary as a function of each study's objectives. The 22-step decision tree was developed so as not to be irrevocably tied to either standard scores or to gains of one-third standard deviation. It is both more general and more permissive than the specific criteria which were adopted for selecting exemplary projects under Contract No. OEC-0-73-6662. It is, in fact, independent of any specific criterion.

Many, if not most of the steps in the decision tree explicitly call for judgments from the evaluator. At each step it is assumed that the evaluator is thoroughly familiar with the issues involved and is qualified to make a judgment based on complex technical considerations. Each decision-tree step is accompanied by a discussion which is intended to define the question that is to be answered, but little or no attempt is made to explain the underlying problems. Such explanations are included in separate appendices in instances where commonly accepted principles or practices are discredited and where new or unusual approaches are endorsed.

It is assumed that the evaluator is familiar with the relevant statistical tools and will apply them appropriately in making his decisions. For this reason, standard statistical procedures are discussed briefly, if at all. More importantly, it should be pointed out that educational evaluation is, and probably will continue to be, an inexact science. Even where the most powerful designs are used, it will be possible to generate plausible hypotheses attributing the observed results to some influence other than the instructional treatment or to factors unique to the tryout site in question. Where weaker designs are employed, it will be highly desirable, or even essential, to strengthen the validity of inferences regarding project effectiveness by amassing as much supporting evidence as possible. In any case, consistency of findings across several replications of an evaluation study would constitute the most convincing kind of supporting evidence.

Figure 1, on pages 46 and 47, summarizes the 22-step decision tree in flow-diagram form. Each step is discussed separately on the pages preceding Figure 1.

The particular path to be followed through the decision tree depends, of course, on the specific design employed in the evaluation study under consideration, but each path is structured so as to focus attention on the design, analysis, and interpretation pitfalls likely to be encountered using that model. Unless a project has been evaluated in several different ways, substantially fewer steps will be required than the 22 which comprise the entire decision tree. Page 2 of Worksheet III, Appendix A was designed for summarizing the considerations of each point in the decision tree and for recording whatever relevant judgments are made.

One additional comment which should be made with respect to the decision tree relates to the fact that it has a number of exit points labeled REJECT. The intent of these exit points is never that the project be rejected as unsuccessful. What is rejected is not the project but the evaluation data which, if the decision-tree process has been carefully followed, have been shown to be inadequate as a basis for reaching any conclusion with respect to the success or failure of the project.

It should be clear from the above and, indeed, from the decision tree itself that exacting compliance with the conventions of experimental design is not generally feasible in real-world educational contexts. Throughout this report the explicit emphasis given to the subjective components of the evaluation process constitutes a deliberate attempt to avoid the misleading impression of algorithmic rigor that might result if the role of judgment were obscured by rigid procedures, arbitrary criteria, and dubious tests of statistical significance.

IV. DECISION TREE FOR VALIDATING STATISTICAL SIGNIFICANCE

Step 1

Question Are the test instruments adequately reliable and valid for the population being considered?

Yes	Proceed to Step 2
No	Reject test scores as measures of project success

Comment The sensitivity of any assessment of instructional impact will be directly related to the reliability and validity of the test instruments used. Evaluation designs which depend on both pre- and posttest scores (e.g., regression models) are especially dependent on highly reliable and valid instruments and, when using such designs, these characteristics should be more heavily weighted in the test selection process than might be appropriate where conventional experimental designs are employed.

Even where conventional experimental designs are used and practical concerns such as testing costs and time will influence instrument selection, reliability and validity considerations must not be ignored. It should also be remembered that the reliability and validity figures cited in test publishers' manuals may not be appropriate for the group being tested or under the circumstances involved. There are several potential problems:

1. The cited reliability coefficients are likely to be measures of internal consistency (e.g., split half, Alpha) rather than measures of temporal

stability (e.g., test-retest). While the two types of reliability estimates tend to be closely related, there may be significant differences, and the concern here is the extent to which the test will yield the same scores on successive administrations.

2. The cited reliability coefficients are likely to be too high if the group to be tested represents only a portion of the grade-level span for which the test is nominally intended.

3. The cited reliability coefficients are likely to be too high if the group to be tested is restricted in its range of ability. Reliabilities for disadvantaged and gifted groups, for example, will be lower than reliabilities for representative groups. A rough reliability estimate for a treatment group with a restricted range of test scores (e.g., bottom quartile) may be obtained from the following formula (Guilford, 1965, p. 464):

$$r_{XX_t} = 1 - \frac{\sigma_n^2 [1 - r_{XX_n}]}{s_t^2}$$

where

r_{XX_t} = reliability for the treatment group

r_{XX_n} = reliability for the norm group

s_t = treatment group pre- or posttest standard deviation (whichever is smaller)

σ_n = norm group standard deviation

> This formula is based on the assumption that the standard error of measurement for the treatment group is equal to the standard error of measurement for the norm group. If the experimental group measurement error is actually higher than that for the norm group, this estimate of test reliability will be too high (see Stanley, 1971, p. 362).

Floor effects will further lower reliability for a group in the lower tail of the distribution, and a judgment must be made as to the impact of these effects (see Ste; 2).

It should be kept in mind that test administration and scoring procedures may have important effects on reliability and validity. Unless the procedures outlined in the publisher's test manual are followed closely, the obtained scores may seriously misrepresent achievement levels. This problem is particularly acute where the effectiveness of an instructional project is assessed by means of norm-group comparisons.

Step 2

Question Are pre- or posttest score distributions of any groups curtailed by ceiling or floor effects?

- Yes — Estimate the size of the effect, record on the worksheet, and proceed to Step 3
- No — Proceed to Step 3

Comment Ideally, the lowest scoring pupil should score above the chance level on the test and the highest scoring pupil should score below the maximum possible score. The actual chance level is difficult to estimate since it depends on the guessing strategy of each student. For students who guessed randomly on all items they didn't know, chance would equal the number of items divided by the number of response alternatives per item. Students often leave items blank, however, even when instructed to guess, and when they do guess, their choices are not necessarily selected randomly from all available alternatives. Because of these problems, the most practical way of identifying floor or ceiling effects is inspection of score distributions for excessive skewness. If the treatment children encounter the test floor on pretesting, or the ceiling on posttesting, their gains will be underestimated. (Gains would only be overestimated where the ceiling was encountered on pretesting and/or the floor on posttesting. This improbable event could occur where different levels of a test were used for pre- and posttesting but there is generally enough overlap between levels so that this type of situation can be avoided.)

If the experimental design employs a control group, it would be subject to similar estimation errors which would

then need to be considered in combination with those of the treatment group.

There is no foolproof method of estimating the size of ceiling or floor effects. In a symmetrical distribution, however, the mean and median will be equal. Compressing one end of the distribution will affect the mean but not the median. The median, then, may provide a reasonable estimate of where the mean would have been in the absence of a ceiling or floor effect.

Step 3

Question Is there reason to believe that the pretesting experience may have been at least partially responsible for the observed treatment effect?

Yes	Estimate the size of the effect, record on the worksheet, and proceed to Step 4
No	Proceed to Step 4

Comment If standardized tests are used, and the experimental design employs a control group, the pretesting experience should have little or no effect on the outcome of the evaluation. Pretesting with criterion-referenced tests, however, may sensitize pupils as to what they are expected to learn. This sensitization may interact differentially with the learning experience available to treatment and control pupils so as to produce greater learning of criterion items in the treatment group.

A more serious problem arises where there is no control group because, as Campbell and Stanley (1963) point out, "...students taking the test for the second time, or taking an alternate form of the test, etc., usually do better than those taking the test for the first time [p. 179]." Since, presumably, children in the norm groups took the test only once, this spurious increment would be present only in the posttest scores of the program participants and could thus lead to erroneous conclusions regarding treatment impact. A compounding of this effect would almost certainly occur if pretesting were the children's first test-taking experience. Under these conditions, pretest scores might be artificially low.

Assuming some test-taking sophistication, a rule-of-thumb estimate for the size of the practice effect would be one tenth of a standard deviation if the same form of the test were used for both pre- and posttesting (Levine & Angoff, 1958.) Use of alternate forms would significantly reduce this effect, but is probably an undesirable practice except in rare cases where matching of the alternate forms is nearly perfect.

Step 4

Question Is there reason to believe that knowledge of group membership may have been at least partially responsible for the observed treatment effect?

Yes Estimate the size of the effect, record on the worksheet, and proceed to Step 5

No Proceed to Step 5.

Comment Knowledge of group membership may produce the Hawthorne effect in members of the treatment group or the "John Henry" effect (Saretsky, 1972) in the control group. [The Hawthorne effect is the occurrence of a performance increment which results, not from the efficacy of a particular treatment, but simply from an awareness that something special is being done. See Whitehead (1938) and Parsons (1974) for further explication. The John Henry effect arises when those who do not receive special treatment make an extra effort in an attempt to demonstrate that they can do just as well without it.] There are other spurious influences of this type which may also confuse the issues. Children may deliberately score poorly on a test in order to get into a special program or to keep from graduating out of a program they enjoy. They may also score poorly to punish a teacher or developer they dislike.

In theory, many of these effects could be experimentally controlled through use of a placebo treatment as is commonly done in medical research. In practice, however, this approach is not feasible, and the educational researcher is left in the unenviable position of having no experimental or statistical technique for controlling such influences. Although they have a tendency to dis-

sipate with time, the researcher has no real recourse but to rely on his own experience and judgment in deciding whether treatment outcomes should be attributed entirely to treatment effects or whether knowledge of group membership increased or decreased the apparent impact. Estimating the size of such effects, of course, can be done only very crudely and even such judgments as "too small to have produced the observed effect" or "large enough to have obscured true project impact" will always be open to question.

Step 5

Question Is there reason to believe that student turnover may have been partially responsible for the observed treatment effect?

Yes	Estimate the size of the effect, record on the worksheet, and proceed to Step 6
No	Proceed to Step 6

Comment Most often, educational evaluations restrict their reporting to include only pupils for whom both pre- and posttest scores are available. While this is the preferred method for dealing with the problem, pupils left out of the analysis because of incomplete data are likely to be systematically different from those included (lower socio-economic status, more mobile families, higher absenteeism rate, higher dropout rate, etc.).

Where pretest and posttest scores are reported on groups which are not identical (i.e., some children have pretest scores only and other have just posttest scores), systematic biases may be present. Students who dropped out, for example, may have been the lower scorers and thus have contributed to a spuriously low mean pretest score and spuriously high apparent gain. Pupils entering a project after it begins may also be atypical and may cause posttest scores to be either too high or low. These possible influences can be checked by comparing pretest scores of the pretest-only group with those of the pre-and-posttest group and by following similar procedures with between-group posttest score comparisons.

Step 6

Question Does the evaluation employ a control group?

Yes Skip to Step 14
No Proceed to Step 7

Comment The term "control group" is used loosely here to connote any comparison group other than a norm group. While the two types of groups serve identical purposes, namely to provide an estimate of how well the treatment group would have performed if it had not received the treatment, normative data generally differ substantially from data collected on control groups, and different analytic procedures must be employed.

Evaluations based on norm-group comparisons are dealt with in the branch of the decision tree which begins with Step 7. Control-group designs are covered in the branch beginning with Step 14.

Step 7

Question — Were pretest scores used to select the treatment group?

Yes	Estimate the size of the regression effect, record on the worksheet, and proceed to Step 8
No	Proceed to Step 8

Comment — It is often the case that children with the greatest educational need are selected for program participation from a larger group of children. If this selection is based on achievement test scores which are subsequently treated as pretest measures, a spurious negative correlation is produced between pretest performance and gains from pre- to posttest. This spurious relationship arises from the fact that scores at the low end of the distribution reflect a preponderance of negative measurement error while those at the high end reflect a preponderance of positive measurement error. Immediate retesting of the extreme groups (using an alternate form of the test) would show the so-called regression effect whereby the mean scores of these groups would move closer to the original total-group mean than they were on the original test.

The magnitude of the regression effect can be approximated by estimating the mean pretest "true" score from the test reliability. To obtain this estimated mean true score for a selected subgroup, the subgroup mean should be subtracted from the total group mean and the difference multiplied by one minus the test-retest or alternate-form (not split-half) reliability. The estimated mean true score is then obtained by adding the result of these calculations to the mean score of the selected subgroup.

It is clear that the size of the regression effect is inversely related to the reliability of the test instrument which is used. For this reason it is important to remember that the reliability coefficients presented in the test publisher's manual are likely to be too high for applications where the group tested represents a restricted range of ability. Step 1 presents a procedure for estimating reliabilities under such circumstances, but it should be noted that even these estimates may be too high and the size of the spurious regression may thus be underestimated.

Step 8

Question Are normative data available for testing dates which can be meaningfully related to the pre- and posttesting of the program pupils?

Yes — Proceed to Step 9

No — Reject norm-group comparisons as adequate evidence of project success

Comment Some test publishers have collected normative data at more than one point during the school year while others have relied on a single data point per year. In either case, it is common practice to publish separate norms tables for the beginning, middle, and end of each school year. Obviously, some of these norms are constructed through processes of interpolation and/or extrapolation. These constructed norms, while possibly useful for counseling or diagnostic purposes, are likely to be in error by amounts large enough to invalidate any inferences drawn about cognitive growth. If they are based on projections of more than a month or two, they should <u>never</u> be used for assessing the impact of educational influences.

Where real (as opposed to constructed) norms are used, they should be treated in the same manner as data from a control group. While even the most naive evaluators would recognize the folly of testing treatment and control groups at significantly different times, test publishers' suggestions that their norms are valid over three- or even four-month periods are rarely questioned. Clearly, however, the treatment group is being compared to a norm group tested at specific times, and unless the testing times of the two groups correspond very closely, any comparisons are

likely to be quite misleading. Ideally, the treatment group should be tested at times exactly corresponding to real normative data points. If this is not possible, linear interpolations or extrapolations of a month or even two months from the specific testing dates on which the norms are based should not introduce large error components. Certainly, it is better to interpolate or extrapolate than simply to use the given norms when the testing times differ. (See also Appendix D.)

Another possibility, where testing times were non-comparable, would be to make explicit the comparisons which were made. An example of this approach might be as follows: "The mean score on the pretest (administered at grade level 7.1) fell at the 24th percentile of the grade 7.5 norm group while the mean score on the posttest (administered at grade level 7.8) was at the 36th percentile of the 8.6 norm group." While this approach may be somewhat confusing, it is scientifically sound whereas other commonly employed approaches (e.g., use of constructed norms) are simply not meaningful.

Step 9

Question Do the norms provide a valid baseline against which to assess the progress of the treatment group?

Yes Proceed to Step 10

No Reject norm-group comparisons as adequate evidence of project success

Comment Ideally, the norm group should be a representative sample of the population from which the treatment group is drawn. Thus, disadvantaged children should be compared against a disadvantaged norm. While some work toward the development of such norms has been accomplished, only nationally representative norms are available for most standardized achievement tests.

It is, unfortunately, necessary to point out that norming practices vary widely from publisher to publisher and that even the best norms may reflect some minor sampling deficiencies. Normative data presented in test publishers' manuals should never be used uncritically without consideration of the total size and representativeness of the norm group.

When groups of disadvantaged children are compared against "national" norms, they are compared against a composite of subgroups, some of which may be like them while others are certainly not (e.g., non-disadvantaged "late bloomers"). For comparisons to be valid, these subgroups must maintain the same relative positions with respect to one another over time, as significant among-group changes would indicate differential group growth rates with respect to the overall norm. At the present time, there is no evidence that different group growth rates occur (despite

the implication of "late blooming"). Thus, while there are potential hazards in using nationally representative norms to assess the progress of atypical groups, it does not appear unreasonable to do so.

Where treatment groups are clearly special (e.g., non-English speaking), national norms should not be assumed to constitute a meaningful basis for assessing progress. One further comment should be made with respect to normative data for grades above the elementary level. Since dropouts come largely from the low end of the distribution, the percentile standing of the non-dropouts will decline. To give an extreme example, if all children below the tenth percentile were to drop out, children originally in the tenth percentile would immediately become first-percentile children. This effect, even in less extreme cases, will cause an apparent negative growth rate among the non-dropouts. Unfortunately, it is not possible to adjust for this phenomenon in the absence of nationally representative empirical data on dropouts.

Step 10

Question Is the comparison between the treatment group and the norm group based on pre- and posttest scores or on gain scores?

Pre- and posttest scores	Proceed to Step 11
Gain scores	Skip to Step 12

Comment Gain scores developed from raw scores or most derived scores are not readily interpretable in norm-referenced evaluations and cannot be interpreted at all in the absence of pretest status information. The problem stems from the fact that the no-treatment expectation in such evaluations is that the group will maintain its percentile standing with respect to the national norm from pre- to posttest. Where pre- and posttest scores are available, it is simpler and less subject to error to work with these measures directly rather than to use gain scores.

Grade-equivalent gains appear to be an exception to this general rule. Gains expressed as grade-equivalent months per month of project exposure seem automatically to provide a comparison with the average child. Not only is this appearance erroneous, but scaling and other problems associated with grade-equivalent gains are so severe that these scores are more misleading than useful (See Appendix D).

Gain scores derived from "regular" standard scores (as opposed to expanded standard scores) constitute the only real exception to the need for pretest scores in norm-referenced evaluations. Where such scores are provided (e.g., for the Gates-MacGinitie) the no-treatment expected gain is 0.0 points. Unfortunately, very few publishers include "regular" standard scores in their test manuals.

Step 11

Question Have appropriate statistical tests been employed to assess the significance of the gain in treatment group performance relative to the norm group?

Yes Skip to Step 22

No Skip to Step 13

Comment The gain of the treatment group with respect to the norm is determined by subtracting the expected mean posttest score from the observed mean posttest score. To find the expected mean posttest score:

1. Determine the percentile equivalent of the mean pretest raw or, preferably, standard, expanded standard, or scale score.
2. Enter the norm table appropriate for the posttest with the pretest percentile and read out the corresponding raw, standard, expanded standard, or scale score (the type of score must correspond to that of the observed mean posttest score). This score reflects the level of performance which would have been expected had there been no special instructional treatment.

The statistical significance of the treatment effect can be assessed using the formula on the following page.

$$t_{N-1} = \frac{\bar{Y} - \hat{\bar{Y}}}{\sqrt{\frac{s_X^2 + s_Y^2 - 2r_{XY}s_Xs_Y}{N-1}}}$$

where $\bar{Y}$ = observed mean posttest score

$\hat{\bar{Y}}$ = expected mean posttest score

s_X = pretest standard deviation

s_Y = posttest standard deviation

r_{XY} = correlation between pre- and posttest scores

N = number of children

$N-1$ = degrees of freedom

Using this formula assumes that normative data are available for testing dates comparable to the pre- and posttest administration times (see Step 8). It is also essential, of course, that the norms be derived from large and representative samples of the treatment group's grade-level peers.

Some test manuals provide simplified procedures for determining the significance of a gain from pre- to posttest. These procedures should <u>not</u> be used, however, as they incorporate assumptions about the correlation between pre- and posttest scores which may not be applicable to the project participants. The significance of the gain should be determined from data in hand.

Step 12

Question Are pre- and/or posttest scores available?

Yes	Proceed to Step 13
No	Reject norm-group comparisons as adequate evidence of project success

Comment Except in those unusual instances where gain scores are derived from "regular" standard scores (scores which have been normalized and standardized independently at each normative data point), it is not possible to derive gain expectations from them. Where gain scores derived from "regular" standard scores are available, the mean gain score can replace the numerator of the formula given in Step 11 and the standard error of the gain (the standard deviation divided by the number of pupils) can replace the denominator of the same equation.

All other gain scores are uninterpretable with respect to expectations. Unless, therefore, it is possible to retrieve pre- and posttest scores, norm-group comparisons cannot provide adequate evidence regarding project success.

Step 13

Question Can appropriate statistical tests be employed to assess the significance of the gain in treatment group performance relative to the norm group?

Yes	Compute appropriate statistics and skip to Step 22
No	Reject norm-group comparisons as adequate evidence of project success

Comment If the mean pretest and posttest scores and the associated standard deviations are available, the statistical significance of the treatment effect can be assessed using the formula given in Step 11, p. 32. If these values are not available and cannot be computed from raw data, norm-group comparisons cannot provide adequate evidence regarding project success.

Step 14

Question — Were the children, either matched or unmatched, randomly assigned to the treatment and comparison groups?

Yes	Skip to Step 18
No	Proceed to Step 15

Comment — A "yes" answer to this question implies that, prior to the beginning of the project, a pool of eligible children existed and each child had an equal chance of being assigned to the treatment group. It further implies that assignment was made on a purely chance basis without any knowledge or consideration of the characteristics of the pupils (except, of course, where matching was done prior to assignment).

If a matching procedure was employed, it should have been implemented as follows. The entire pool of eligible children should have been organized into carefully matched pairs on the basis of pretest scores and other potentially relevant variables (e.g., sex). One member of each pair should then have been selected at random for assignment to the treatment group. The remaining member of the pair would then, of course, have been assigned to the comparison group.

Note: Matching after assignment to treatment and comparison groups is a fundamentally unsound practice. (See Step 15.)

Step 15

Question — Is there evidence that members of the treatment and control groups belong to the same population or to populations that are similar on all educationally relevant variables including pretest scores?

Yes	Proceed to Step 16
No	See Appendix C

Comment — Random assignment will usually (but not always) produce groups which are comparable. On the other hand, groups resulting from non-random processes are likely to differ from one another on educationally relevant dimensions. If such differences exist, there is no entirely satisfactory means of making between-group comparisons.

As Lord (1967) has pointed out, "If the individuals are not assigned to the treatments at random, then it is not too helpful to demonstrate statistically that the groups after treatment show more difference than would have been expected from random assignment—unless, of course, the experimenter has special information showing that the nonrandom assignment was nevertheless random in effect [p. 38]." The same could be said where significant pretest differences were found between groups which were developed through random processes.

Where pre-existing, intact groups are used as the treatment and control groups, it is not appropriate to assume that they are, even in effect, random samples from a single population. The probability that they may be must be investigated empirically. At the very least,

the two groups must not be significantly different in terms of pretest scores. They should also be comparable in terms of socioeconomic status, age, sex, and racial and ethnic composition. School size and setting (urban - rural) as well as neighborhood should also be comparable. Even with these factors equated, serious selection biases are common. Such biases are introduced when teacher or student participation is voluntary or when experimental groups are selected by principals or teachers.

A common design error where comparable, intact groups cannot be found is that of matching members of the treatment group with specific members of other, non-comparable groups. The assumption here is that a comparable control group can be constructed through the matching process. The fallacy inherent in this assumption is that the selected subgroup is atypical of the group from which it is drawn and will show a regression toward the mean of that group on posttest measures. Campbell and Stanley (1963) describe this type of post-hoc matching as "a stubborn, misleading tradition in educational experimentation," and as a "hazard" which is "frequently tripped over [p. 219]."

Step 16

Question — Are post-treatment comparisons made in terms of posttest or gain scores?

Posttest scores	Skip to Step 19
Gain scores	Proceed to Step 17

Comment — Two types of gain score are frequently used in educational evaluations: raw and residual gain scores. Raw gain scores are derived by subtracting pretest scores from posttest scores. When raw gain scores are used, the size of the treatment effect is defined as the treatment group's raw gain score minus that of the control group. It can be shown that this difference is mathematically identical to the treatment group's posttest score minus the control group's posttest score after the latter has been adjusted by the entire amount of the difference between the two groups' pretest scores. Compared to covariance analysis, which the authors hold to be the most appropriate method to compensate for initial differences between groups, the raw gain score adjustment is excessive and results in an overestimation of the treatment effect when the treatment group's pretest score is lower than that of the control group. Conversely, raw gain scores underestimate the size of the treatment effect when the treatment group scores higher on the pretest than the control group.

Residual gain scores are not really gain scores at all. They are differences between observed posttest scores and posttest scores predicted from the regression of posttest on pretest scores for the combined treatment

and control groups. If the treatment has been effective, observed performance of the treatment group on the post-test will exceed the prediction, whereas the performance of the control group will fall below the predicted value. The sum of the absolute values of the two deviations is presumed to yield a measure of the treatment effect. Where there is no difference between groups on the pre-test, covariance analysis, raw gain scores, residual gain scores, and even simple posttest comparisons will all yield exactly the same measure of the treatment effect. As pretest differences are introduced, however, the measure of treatment effect obtained from residual gain scores systematically diminishes and approaches zero where initial between-group differences are large. Where any pretest differences exist, a residual gain analysis will always underestimate the size of the treatment effect.

Wherever possible, covariance analysis, preferably with an adjustment for test unreliability (e.g., Porter, 1967), should be used to compensate for initial differences between treatment and control groups, — assuming, of course that the two groups can be regarded as random samples from a single population. Statistically significant treatment effects found with either residual gain scores or raw gain scores when the treatment group is initially inferior to the control group constitute adequate evidence of project success. The real danger inherent in these approaches lies in the rather high probability of rejecting projects which are really effective.

Step 17

Question Can data be obtained which would enable application of covariance analysis techniques, would such analyses be appropriate, and is there a reasonable expectation that they would produce significant results?

Yes	Conduct covariance analysis and proceed to Step 22
No	Skip to Step 20

Comment Wherever pretest differences between treatment and control groups have resulted from random assignment procedures, covariance analysis may be employed to adjust for these differences. Where the treatment group was superior on the pretest, this type of analysis will significantly reduce the probability of incorrectly inferring a treatment was successful when it was not. Conversely, where the treatment group was initially inferior, covariance analysis will significantly reduce the probability of rejecting a successful treatment as unsuccessful. In both instances the covariance adjustment will increase the accuracy of posttest measures so that the true magnitude of program impact can be determined.

There is, of course, no justification for the extra computational labor required for covariance analysis if the two groups obtained equal scores on the pretest. Further, covariance analysis is not required where an initially inferior treatment group scored significantly higher than the control group on the posttest if interest is restricted to the statistical significance of the treatment effect rather than an estimate of its size.

Step 18

Question Were pretest scores collected?

Yes	Go back to Step 15
No	Proceed to Step 20

Comment If assignment of pupils to treatment and control groups has been truly random, it is not essential to collect pretest scores since valid inferences can be drawn from posttest score comparisons. If pretest scores are collected, however, more powerful statistical tests can be employed in cases where the assignment process has resulted in small initial differences between the groups.

Step 19

Question Have covariance analysis techniques been employed to adjust for initial differences between groups?

Yes	Skip to Step 22
No	Go back to Step 17

Comment Where assignment to either the treatment or the control group has been random or "random in effect" (see Step 15), small pretest score differences may be found between groups. Under these circumstances, analysis of covariance is the most appropriate statistical technique available for testing treatment effects. If the analysis has been done correctly, its findings may be accepted at face value.

Covariance analysis must never be regarded as an adequate technique for statistically equating dissimilar groups. It can only be used where its assumptions (effectively random assignment and homogeneity of regression) are met and where initial differences between groups are not excessive. It should be noted that even where regression is statistically non-heterogeneous, small differences in regression line slopes introduce errors into the computations. These errors interact in a multiplicative fashion with the size of the between-group difference. A small error multiplied by a big difference becomes a big error. For this reason, it is common to use the 10% level for rejecting the hypothesis of homogeneous variance. Use of the 20% level would be appropriate when the difference between group means is large.

Step 20

Question Have appropriate statistical tests been employed to compare posttest or gain scores?

Yes	Skip to Step 22
No	Proceed to Step 21

Comment A wide variety of statistical tests and procedures can be used for testing differences between groups. Raw or (preferably) standard score comparisons may often be made on either posttest or gain scores using parametric statistical tests such as Student's *t* for independent means (*t* for correlated scores where pupils were matched prior to assignment to groups) or analysis of variance. However, the data should be inspected to confirm that the assumptions of these tests have been met, since score distributions from special instructional projects may not meet requirements such as normality due to test ceiling or floor effects or other confounding influences.

Where parametric test assumptions are not met, non-parametric tests such as the Mann-Whitney *U* or the Kolmogorov-Smirnov test are appropriate but are less powerful than their parametric equivalents. Non-parametric tests must also be used where comparisons are made between posttest grade-equivalent scores (assuming random assignment). There is no meaningful way in which grade-equivalent gains can be compared.

The cautions regarding the drawing of inferences from gain-score comparisons discussed in Step 16 should be carefully observed.

Step 21

Question Can data be obtained which would enable appropriate tests to be made?

Yes	Obtain data, compute appropriate statistics, and proceed to Step 22
No	Reject posttest and/or gain score comparisons as adequate evidence of project success

Comment Where inappropriate statistical approaches have been adopted, there is no choice but to seek out the information needed to conduct appropriate tests. If raw or (preferably) standard score summary statistics (means and standard deviations) are available, *t*-tests could be done. In many cases, unfortunately, all calculations will have been done inappropriately (e.g., by using grade-equivalent scores) and it will be necessary to go back to individual test scores if meaningful analyses are to be done. If this procedure is followed, raw or grade-equivalent scores should be converted to their standard-score equivalents before any arithmetic operations are performed on them. Appropriate tests are discussed in Steps 17 and 20.

Step 22

Question Do analysis results favor the treatment group at the preselected level of statistical significance?

- **Yes** Review all evidence compiled during the validation process and use judgment to decide whether the statistical test results can reasonably be attributed to project effects
- **No** Reject evidence as being inadequate to validate project success

Comment Given a statistically significant result, the attribution of cause is still at issue. The final step in relating an observed effect to the treatment requires careful consideration of each of the extraneous effects identified in proceeding through the decision tree and estimation of their contribution, in aggregate, to the apparent impact of the treatment. It is, finally, left to the judgment of the evaluator to assess the magnitudes of these effects, weigh their influence in the evaluation results, and conclude whether or not the treatment was effective.

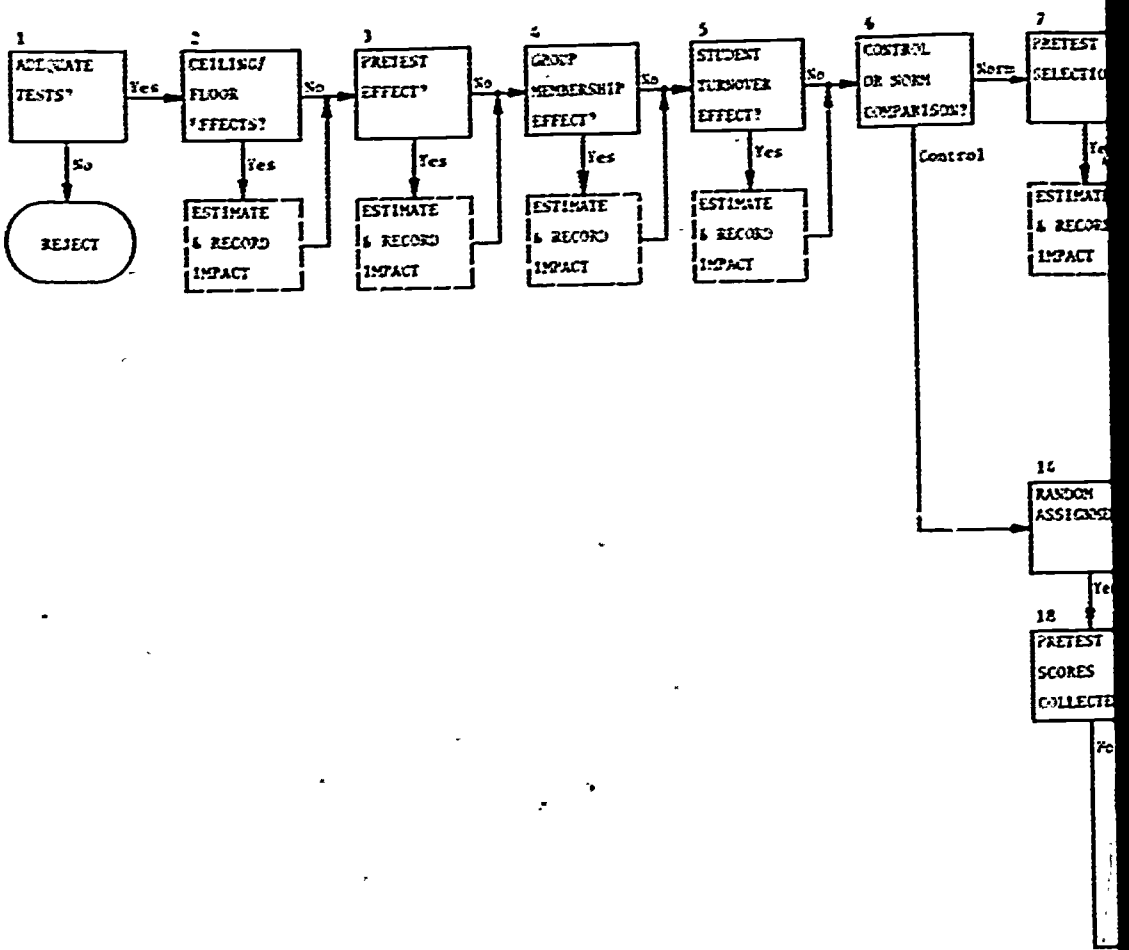
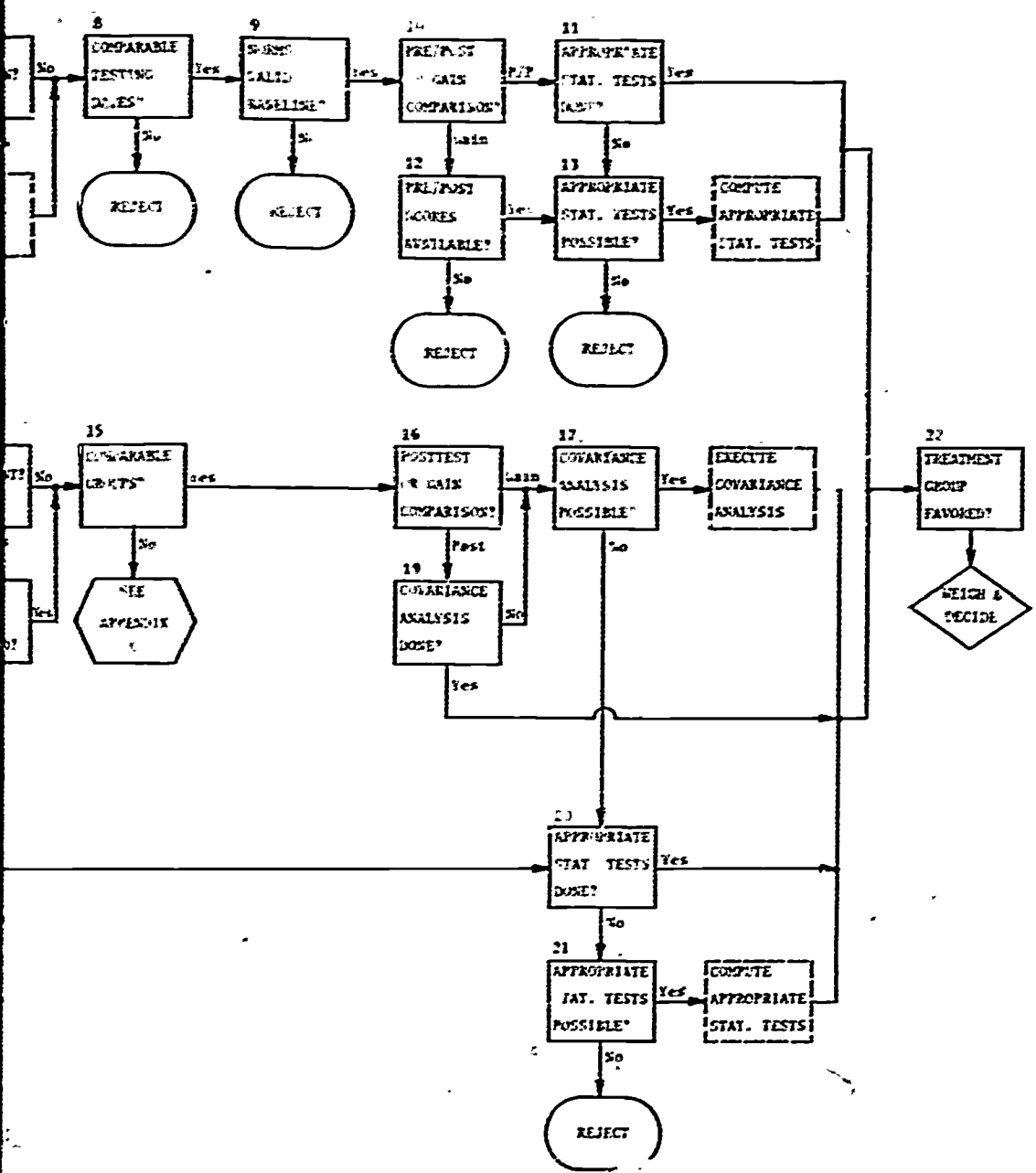


Figure 1. Decision tree for validating statistical significance.


8
COMPARABLE TESTING DATES?
Yes
No
9
SCORES VALID BASELINE?
Yes
No
10
PRE/POST GAIN COMPARISON?
P/P
Gain
11
APPROPRIATE STAT. TESTS (ANY)?
Yes
No
12
PRE/POST SCORES AVAILABLE?
Yes
No
13
APPROPRIATE STAT. TESTS POSSIBLE?
Yes
No
COMPUTE APPROPRIATE STAT. TESTS
REJECT
REJECT
REJECT
REJECT
15
COMPARABLE UNITS?
Yes
No
SEE APPENDIX
16
POSTTEST OR GAIN COMPARISON?
Gain
Post
17
COVARIANCE ANALYSIS POSSIBLE?
Yes
No
EXECUTE COVARIANCE ANALYSIS
19
COVARIANCE ANALYSIS DONE?
Yes
No
22
TREATMENT GROUP FAVORED?
WEIGH & DECIDE
20
APPROPRIATE STAT TESTS DONE?
Yes
No
21
APPROPRIATE STAT. TESTS POSSIBLE?
Yes
No
COMPUTE APPROPRIATE STAT. TESTS
REJECT

V. ADDITIONAL CONSIDERATIONS

The decision tree presented in the preceding section of this report should enable reasonably unequivocal conclusions to be reached regarding the existence or nonexistence of some treatment impact. Difficult as that decision-making process may be, even more difficult questions arise in assessing the practical value of the observed impact. Relevant questions include, "What is the educational significance of a third-of-a-standard-deviation (or any other size) gain on a standardized reading achievement test?", "What is the significance of a five-point gain in reading comprehension as opposed to a comparable gain in vocabulary?", and "Is a moderate-cost treatment which produces moderate gains more educationally significant than a costly treatment which produces larger gains?"

Consideration of these and related questions quickly brings to light the difficulty of making even gross-level decisions in the absence of a metric for quantifying educational significance. And many would argue that scores on standardized achievement tests in no way satisfy the requirements for such a metric. Unfortunately, the lack of a presumably adequate metric for educational significance does not relieve decision-makers of their responsibility to choose among and act upon the alternatives available to them. Neither does the lack of an adequate metric imply that all measurement is infeasible or that decisions must be made without useful guidance from educational research. Standardized test scores do constitute meaningful indices and, if appropriately interpreted, go a long way toward achieving their ultimate objective.

Basic to the entire quantification issue is the sometimes overlooked fact that educational significance is an inherently subjective concept. While scales may be constructed from the consensus of experts, it must be acknowledged that they will be culture-bound and situation-specific.

Furthermore, there will be educators of substantial stature who will disagree with any set of consensus-based priorities and relationships.

A simple illustration can be drawn from standardized reading achievement tests where it is common practice to provide separate scales for vocabulary, comprehension, and occasionally other component skills. Clearly these subtests could be weighted and combined in a number of different ways to yield a "Total Reading" score. Some educators might argue that vocabulary and comprehension are equally important aspects of reading while others might claim that comprehension was twice—or five times—or even ten times as important as vocabulary. It is clear that this issue cannot be adequately resolved through empirical research and can only be dealt with by "majority rule" or some similar, equally unsatisfactory expedient.

Despite the fervor with which this issue may be debated, the method of combining vocabulary and comprehension subtest scores to obtain a total reading score appears, upon closer examination, to be little more than a pseudo-problem. The two subtests are so highly intercorrelated (typically, $r = .80$) that even very different weighting systems have almost no impact on the ordering of total scores. In other words, students will fall into very nearly the same order whether comprehension scores are given ten times the weight of vocabulary scores or the two scales are equally weighted. Although the empirical evidence may be less complete, it appears that many widely debated issues in educational evaluation today can be deflated with the same sort of demonstration. Clearly, the argument that standardized achievement tests ought not to be used for assessing cognitive growth can be quickly invalidated if the correlations between test scores and other measures purported to reflect component skills more adequately are shown to be high.

The conclusion, then, must be that standardized tests, with all their deficiencies, do provide a useful metric for assessing the basic skills of reading and math. Standard scores on such tests, although not comprising ratio scales, do provide a means of quantifying gains,

of relating observed gains to gain expectations in a reasonable manner, and of measuring the impact of special instructional projects on cognitive growth. At the same time, it is clear that they do not provide a complete answer to the kinds of questions raised in the first paragraph of this section. The difficulty in coming to grips with these questions lies not in determining the size of the gains but in determining their value.

The value issue was alluded to above in discussing the relative value of gains in vocabulary as opposed to comprehension. In this situation, at least, the issue was shown to be a pseudo-problem and it was implied that many similar issues might be of far greater theoretical than practical concern. The absolute value of achievement gains may also pale into relative insignificance when examined in the context of real-world contingencies. An achievement gain of "X" standard-score points is likely to be worth exactly the amount of money a school district is able or willing to spend to obtain it--and this, in turn, will depend on the needs of the children in the district and perceptions of the relative priorities existing among them. If needs can be adequately defined, relative comparisons among the alternatives available to fit them are sufficient. Absolute scales of educational significance may be required for the typical kind of cost-benefit studies seen in the harder science and engineering areas, but educational issues need not be defined in that manner.

In their search for effective compensatory education projects to package, the authors decided they would consider any treatment which produced one-third of a standard deviation gain with respect to the national norm. Above that point, choices would be based on judgments reflecting the size of gains, costs, replicability, availability, target group served, variety of approach, etc. Their original guess that the choices would be relatively easy to make and unequivocal was substantiated. While this example may be atypical, it seems that the alternatives available to fill a specific need will rarely be so numerous as to preclude sound decision-making by qualified, well-informed, and thoughtful judges.

APPENDIX A

PROJECT SELECTION CRITERIA WORKSHEET I

SUMMARY PAGE

PROJECT TITLE _____

Date	Initials	
		DESCRIPTION Approach Pull-out vs. Whole class
		PREREQUISITES ☐ Provides instruction in reading and/or math ☐ Serves children in grades K-12 ☐ Serves educationally disadvantaged children ☐ Has achievement test data for more than one "instance"
		FINAL ASSESSMENT ☐ Accepted ☐ Rejected Reason for rejection ☐ Prerequisites not met ☐ Inadequate evidence of effectiveness ☐ Excessive costs ☐ Not available ☐ Not replicable

PROJECT SELECTION CRITERIA WORKSHEET II

PRELIMINARY SCREENING CRITERIA

AVAILABILITY

Accessibility:

- ☐ Can be visited for validation
- ☐ Personnel are cooperative
- ☐ Procedures, results, and costs are documented

Acceptability:

- ☐ Operational in public schools
- ☐ Not primarily a single commercial product

COST

- ☐ Equipment plus special personnel less than $___ per pupil
- ☐ Initial investment less than $___ per pupil
- ☐ (Alternatively) Per-pupil cost over a three year operational period including start-up costs should not exceed $___ per year

REPLICABILITY

- ☐ All major components can clearly be duplicated. Components include: materials, hardware, personnel, and environments.

EFFECTIVENESS

- ☐ Achievement test data show consistently that actual post-treatment performance exceeds the no-treatment expectation by an amount which is statistically significant and equal to at least _____ standard deviation with respect to the national norm.

PROJECT SELECTION CRITERIA WORKSHEET III

ANALYSIS OF PROJECT EVALUATION

Complete a separate sheet for each validating site or combination of sites for which separate data are reported.

PROJECT TITLE _____

Tryout Group _____

I. Tryout Summary

A. Treatment group description

1. Number _____
2. Grades/Ages _____
3. SES/Ethnic _____
4. Pre-project achievement level _____
5. Schools/Classrooms _____
6. Selection procedure _____
7. Treatment period dates _____

Hours per week _____

B. Comparison group description (if same as experimental group write "same")

1. Number _____
2. Grades/Ages _____
3. SES/Ethnic _____
4. Pre-project achievement level _____
5. Schools/Classrooms _____
6. Selection procedure _____
7. Treatment period dates _____

Hours per week _____

PROJECT SELECTION CRITERIA WORKSHEET III (Continued)

ANALYSIS OF PROJECT EVALUATION

II. Evaluation Model Employed

- ☐ Norm-referenced
- ☐ Control group
- ☐ Regression
- ☐ Other (specify) _____

III. Confounding Influences (comment on items checked)

- ☐ Inadequate tests _____
- ☐ Ceiling/Floor effects _____
- ☐ Pretest effect _____
- ☐ Group membership effect _____
- ☐ Student turnover effect _____
- ☐ Inappropriate testing times _____
- ☐ Inappropriate comparison group _____
- ☐ Participant selection via pretest _____

IV. Evaluation Outcomes

A. Evidence of Statistical Significance _____

B. Size of Gain with Respect to the National Norm _____

APPENDIX B

Norm-referenced versus Criterion-referenced Tests

While use of criterion-referenced tests has been advocated for at least ten years (Glaser & Klaus, 1962), educational projects are still evaluated predominantly in terms of commercial, norm-referenced tests. The reluctance of educators to abandon familiar testing paradigms is understandable in view of the continuing confusion over the exact distinction between the conventional norm-referenced test and the new criterion-referenced instruments. This confusion is clearly evident in recent articles by Airasian and Madaus (1972), Jackson (1971), and Popham and Husek (1971), and in a review by Davis (1973) of eight 1972 AERA papers on criterion-referenced testing.

The confusion appears to result from conceptualizing criterion-referenced tests as an alternative to norm-referenced tests. In fact, norm- and criterion-referenced tests do not represent mutually exclusive test categories nor do they represent the ends of a continuum. On the contrary, the "norm" and "criterion" descriptors refer to completely independent test characteristics, both of which should probably be included in the description of any test. The problem is further complicated by the fact that, although there are real differences between tests that are labeled "norm-referenced" and those labeled "criterion-referenced," these labels do not capture the salient distinguishing features.

The dominant characteristic of tests that are labeled "criterion-referenced" is that their content is clearly defined in terms of some performance dimension of interest. This relationship permits direct interpretation of individual scores in ways which have immediate practical implications (e.g., time required to run a mile, or proportion of the 3000 most common English words that the individual can define). The misleading label apparently derives from the failure to distinguish

between the dimension being measured and the scale adopted to measure it. This failure is not surprising in the context of training program development which first popularized "criterion-referenced" testing. For example, Glaser and Klaus (1962) wrote:

> Two kinds of criterion standards are available for evaluating individual proficiency. First, a standard can be established which reflects the minimum level of performance which permits operation of the system....At the other extreme, proficiency can be defined in terms of maximum system output. The standard of measurement is then expressed as a function of the capabilities of other components in the system. The man loading a Navy gun, for example, never needs to load more rapidly than he receives shells from the magazine below decks. In this case, a fairly absolute standard of proficiency is available. [p. 424]

In this and similar situations, it has become popular to say that a performance criterion has been established and the test used in measuring performance need only tell us whether or not the criterion is reached. It might be more informative to say that the test measures a performance dimension (speed of loading), that system requirements dictate a specific cutoff score, and that in the interest of economy it would be adequate to dichotomize the speed of loading scale about this cutoff. Everyone below the cutoff would get a score of "too slow." Everyone above the cutoff would get a score of "fast enough."

The term "norm-referenced" has rivaled "criterion-referenced" in terms of confusion generated. Any test becomes a norm-referenced test as soon as a norm group of one or more entities is defined and scores of those entities are obtained. Of course, if the norm reference is to be of any use there are many properties that the test and the norm group must have. The required properties depend entirely on the intended use of the test, but one typically desires relevance and proper sampling for norm groups, while tests should provide reliable and efficient quantification.

The relative independence of norm referencing and performance referencing can be illustrated by an instrument used to select students for pilot training. Successful tests for this purpose can and have been

developed using what are usually referred to as conventional norm-referenced test development procedures. It should be clear from the above discussion, however, that norm reference is not the salient characteristic of such tests. While validation groups must be used to develop and scale the tests, the ultimate criterion is flying success, and is not dependent on standings in relation to any norm group. Once a reliable test has been developed which correlates highly with a measure of pilot success, a single cutoff score, or criterion, could be determined, and applicants could be scored either pass or fail.

At the same time, neither the procedures for developing the test nor the final appearance of the test would classify it as "criterion-referenced." That is, it is unlikely that the population of pilot skills would be sampled at all. Of course, one could say that the final instrument defined something called "pilot aptitude" but it is doubtful whether the concept could be identified from the test items or that one would feel enlightened to know that a person who scores "X" or more points on this aptitude could be taught to fly. An "aptitude" as measured by correlated items is simply not what we usually mean by a performance dimension. In short, this most familiar type of test is neither particularly "norm-referenced" nor particularly "criterion-referenced."

It should be noted that the concepts discussed above are not new and have been recognized by various authors (e.g., Glaser & Nitko, 1971; Davis, 1972). Even these authors, however, preserve the norm/criterion-reference categories. Regardless of the terminology which is ultimately adopted, it must be recognized that new and useful measurement techniques have been introduced in the process of attempting to define and develop criterion-referenced tests. It should be emphasized that it is the categorization that is aproductive, and not necessarily the techniques which have been developed.

Implications for Project Evaluation

In contrast to the pilot-trainee selection test which was neither norm- nor "performance"-referenced, the commercial reading and math achievement tests used in project evaluation are both norm referenced and performance referenced. The norm group properties need little comment except to point out that norm groups are typically presented as nationally representative (although some are clearly more representative than others) and may not be suitable for assessing the gains of particular subgroups.

The performance dimension that is defined by standardized tests is somewhat arbitrary, and it may well be argued that substantial improvement is needed here. Raw scores are seldom reported in a meaningful way and items are probably chosen on the basis of discrimination rather than as a sample of a carefully defined performance domain. The problems are almost certainly worse in testing reading than in testing math, but they reflect the basic difficulty in defining what is meant by reading skill and measuring it.

While commercial standardized tests are clearly not optimal instruments for research purposes, there is little empirical evidence to suggest that tests developed according to criterion-referenced procedures provide better measures of project effectiveness in basic skill areas. While, in theory, criterion-referenced instruments which are focused on the specific objectives of a particular instructional treatment ought to be more sensitive to achievement gains resulting from it than the more general standardized tests, the latter clearly sample important aspects of reading and math achievement and are relatively efficient and reliable instruments. Clearly, criterion-referenced or other special-purpose tests are perfectly acceptable for use in assessing the statistical significance of project impact. If enough is known about their properties, it should also be possible to estimate the educational significance of observed gains. One requirement, of course, is that both the statistical and educational significance of

pre-to-posttest gains must be assessed against the gains which would be expected under no-treatment conditions. In the absence of normative data, the estimation of no-treatment posttest status clearly necessitates the use of a comparison group evaluation model.

APPENDIX C

Estimation of Treatment Effects from the Performance of Non-comparable Control Groups

Where treatment and control groups are significantly different from one another, it is generally not possible to assess the impact of an educational intervention. In the case where a treatment group scores lower on the pretest and higher on the posttest than an otherwise comparable control group, it is probably safe to conclude that the treatment was effective but, even here, the magnitude of the treatment effect cannot be accurately estimated.

There are some evaluation designs which employ a non-comparable control group to generate an estimate of how the treatment group would have performed on the posttest had they not participated in the treatment. The most widely applicable and plausible of these designs require that an original group be dichotomized about some pretest cutoff score so that all pupils scoring on one side of the cutoff score receive the treatment while none of those scoring on the other side are allowed to participate. Two such designs are presented here along with one design which does not require such dichotomization. The designs are:

A. The Regression-discontinuity Model
B. The Regression Projection Model
C. The Generalized Multiple-regression Model

A. The Regression-discontinuity Model

The model which appears most immune to plausible alternative hypotheses is the Regression-discontinuity Model (Campbell & Stanley, 1963). A comprehensive development of this model and related statistical tests is available (Sween, 1971). The model requires that treatment and comparison groups be developed from a single original group by assigning all members on one side of a pretest cutoff score to the treatment group and all members on the other side to the comparison group. Separate

pretest-posttest regression lines are then computed for each group and the difference between the lines is tested at the point where they intersect the pretest cutoff value.

The model is rigorous in the sense that, if the procedures are followed correctly, rejection of the null hypothesis for any reason other than a treatment effect is extremely implausible. There are two considerations, however, which severely restrict the applicability of the model. First, it is difficult in a school environment to enforce assignment to treatment groups solely on the basis of test scores, or even on the basis of scores reflecting both test performance and a numerical teacher rating. Second, the model is not sensitive to changes in regression line slopes unless these changes are accompanied by a discontinuity of the regression lines. This requirement represents a potential problem since compensatory education projects are often individualized on the basis of student need. Such individualization could produce the greatest improvement in those students farthest below the pretest cutoff score thereby flattening the treatment-group regression line without producing a discontinuity at the cutoff point. At least one compensatory reading project known to the authors appears to produce this kind of effect.

In short, regression-discontinuity analysis is recommended for all cases in which the conditions for its implementation are met and a positive result can be anticipated. It seems unlikely, however, that such cases will occur frequently.

B. The Regression Projection Model

The Regression Projection Model uses a regression line calculated from the comparison-group pretest-posttest distribution to estimate what the treatment-group posttest scores would have been under a "no treatment" condition. Like the Regression-discontinuity Model, it also requires dichotomization of a total group into treatment and comparison subgroups about a particular pretest cutoff score. The advantage of this model is its sensitivity to treatment-produced changes in regression line slopes. Its primary weakness is its inability to distinguish treatment

effects from other factors which may affect the regression line.

The model is analogous to the technique of Karl Pearson for estimating total-group test validity when criterion measures are available only for those who score above some selected cutoff point. It is applicable where selection (pretest) scores are available for an entire group, but where there is no indication of how the subgroup below the cutoff score would have done on the posttest had they been treated in the same manner as the group above the cutoff.

The basic assumption of the model is that under no-treatment conditions the regression of posttest scores on pretest scores for the total group would be homogeneous and linear throughout the entire score range. The regression line for the comparison group is taken as the estimate of this total group regression line, and is projected through the treatment-group distribution (See Figure 2). This projected regression line is then used to calculate the estimated no-treatment posttest score.

The model should be applied with caution since the basic assumption of homogeneous, linear regression may not be tenable. For example, in compensatory projects, factors which lower the pretest-posttest correlation for low-scoring students may invalidate the model completely. Floor effects on the pretest and other factors leading to low pretest reliability at the lower end of the range are particularly troublesome. At a minimum, a good argument that such factors are not acting is required. A scatter diagram permitting inspection of the pretest-posttest distribution for irregularities is essential.

Horst (1966), Chapter 26, provides a discussion of the underlying statistical issues and presents formulas for generating unbiased estimates of the mean, standard deviation, and pretest-posttest correlation for the total group. The estimated regression equation for the total group is identical to the regression equation for the restricted (comparison) group. Thus, one needs only to calculate the regression equation for the comparison group and use it to obtain estimated treatment-group posttest scores. This equation can be written:

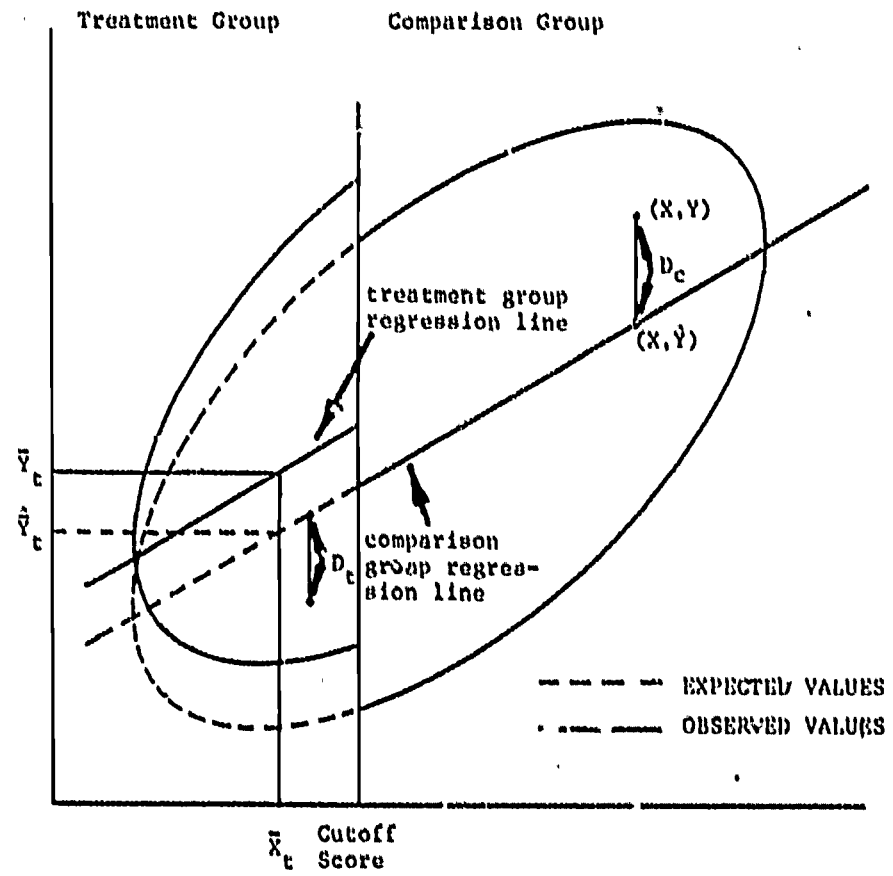


Figure 2. Regression Projection Model

$$\hat{\bar{Y}}_t = b_c \bar{X}_t + k_c$$

where b_c is the slope of the comparison-group regression line and k_c is its Y-axis intercept.

If the mean pretest score of the treatment group is substituted for $\bar{X}_t$ in the above equation, $\hat{\bar{Y}}_t$ will be the estimated mean posttest score ($\hat{\bar{Y}}_t$). The difference between the actual and estimated posttest scores can then be tested using

$$t_{N-3} = \sqrt{\frac{P_t^2(\bar{Y}_t - \hat{\bar{Y}}_t)^2(N-3)}{b_c^2 \bar{s}_X^2 + \bar{s}_Y^2 - 2b_c \bar{b} \bar{s}_X^2 + P_t P_c (\bar{Y}_t - \hat{\bar{Y}}_t)^2}}$$

where P_t = proportion of pupils in the treatment group

P_c = proportion of pupils in the comparison group

N = number of pupils in the combined group

$\bar{s}_Y^2$ = weighted mean of the treatment- and comparison-group posttest variances

$\bar{s}_X^2$ = weighted mean of the treatment- and comparison-group pretest variances

b_c = slope of the comparison-group regression line

$\bar{b}$ = weighted mean of the slopes of the treatment- and comparison-group regression lines

The derivation of this test is not available in the literature and is sketched in its entirety below. Readers not interested in this derivation should skip to the discussion of the Generalized Multiple-regression Model which begins on page 71.

Significance Test for the Regression Projection Model[1]

Consider first the general situation in which a regression line is fit to a pretest-posttest score distribution, providing an estimated

1. We are grateful to Paul Horst for the rationale and development of this test. However, the authors are responsible for the presentation given here and for any errors it may contain.

posttest score ($\hat{Y}$) for each pretest score (X). The equation for the regression line may be written

$$\hat{Y} = bX + k$$

where b = slope of the regression line

k = Y-intercept of the regression line

Then, for each student, we can define a value

$$D = Y - \hat{Y}$$

which is the difference between his actual posttest score and his estimated posttest score or, in other words, the distance that his actual posttest score is above or below the regression line.

Next, consider the Regression Projection Model in which a regression line is fit to the comparison-group data and then projected through the treatment-group data (Figure 2). A distance D_c from this regression line can be computed for each comparison-group student. A distance D_t from the same comparison-group regression line can be computed for each treatment-group student. Because the regression line was fit to the comparison-group data, the mean of the comparison-group D values ($\bar{D}_c$) will be zero. However, the mean of the treatment-group D values ($\bar{D}_t$) will not be zero unless the mean of the treatment-group posttest scores falls exactly on the projected regression line, that is unless $\bar{Y}_t = \hat{\bar{Y}}_t$.

The null hypothesis which is tested in the Regression Projection Model includes three major conditions: (a) students are assigned to treatment and comparison conditions solely on the basis of their pretest (either single or composite) scores, (b) posttest on pretest regression is linear throughout the range of pretest scores, and (c) there is no treatment effect. If it can be assumed that the first two conditions are met, and if there is no treatment effect, the regression lines of the treatment group, the comparison group, and the total group should all approximately coincide. Deviations of treatment-group posttest scores from the projected comparison-group regression line would have an expected mean value of zero under these conditions so that a sizeable

departure from this expectation may indicate a significant treatment effect. In an experimental situation, we can test whether the observed mean deviation ($\bar{D}$) is larger than would be expected under the conditions of the null hypothesis by computing

$$t = \frac{\bar{D}}{s_{\bar{D}}} . \tag{1}$$

On page 64, t is expressed as a function of treatment- and comparison group statistics. The equation is derived as follows:

First we recall that

$$s_{\bar{D}} = \sqrt{s_D^2/df_D} . \tag{2}$$

Substituting (2) into (1) we may write (1) as

$$t^2 = \frac{\bar{D}^2(df_D)}{s_D^2} \tag{3}$$

We can then develop the numerator and denominator of (3) separately:

Numerator

The combined mean of the D values can be expressed in terms of the mean D values for the two groups (with respect to the comparison-group regression line) and the proportions of cases in each group:

$$\bar{D} = P_t\bar{D}_t + P_c\bar{D}_c \tag{4}$$

But since the regression line was fit to the comparison-group data,

$$\bar{D}_c = 0, \tag{5}$$

Substituting (5) into (4):

$$\bar{D} = P_t\bar{D}_t. \tag{6}$$

And since the mean of the D values is equal to the difference between the means of the observed posttest distribution and the estimated posttest distribution, we can rewrite (6) as:

$$\bar{D} = P_t(\bar{Y}_t - \hat{\bar{Y}}_t). \tag{7}$$

The remaining factor in the numerator of (3) is df_D, the number of degrees of freedom for the standard deviation of D. Usually df_D is taken to be N-1 where N is the number of pairs of observations. However, two additional restrictions hold in this model. First, the comparison-group D values must sum to zero and second, the mean of the estimated posttest scores for the treatment group is determined by the comparison group data. Therefore

$$df_D = N - 3. \tag{8}$$

By combining (7) and (8), the numerator of (3) can finally be written

$$\bar{D}^2(df_D) = \left\{P_t(\bar{Y}_t - \hat{\bar{Y}}_t)\right\}^2 (N - 3). \tag{9}$$

Denominator

It is well known that the variance of a difference between paired measures is equal to the sum of the variances of the two measures minus a correction for the correlation between them. In the case of D values from the Regression Projection Model,

$$s_D^2 = s_{\hat{Y}}^2 + s_Y^2 - 2r_{\hat{Y}Y}s_{\hat{Y}}s_Y \tag{10}$$

where

$r_{\hat{Y}Y}$ = the correlation between actual and estimated posttest scores

s_Y = the standard deviation of the actual posttest scores

$s_{\hat{Y}}$ = the standard deviation of the estimated posttest scores.

Since, by definition,

$$\hat{Y} = b_c X + k_c \tag{11}$$

it can be readily shown that

$$s_{\hat{Y}} = b_c s_X \tag{12}$$

and

$$r_{\hat{Y}Y} = r_{XY} \tag{13}$$

where r_{XY} is the pretest-posttest correlation for the combined group.

Therefore, substituting (12) and (13) in (10)

$$s_D^2 = (b_c s_X)^2 + s_Y^2 - 2b_c r_{XY} s_X s_Y \tag{14}$$

This form of the denominator could be used for computing t. However, since the treatment and comparison groups are normally analyzed separately, it is desirable to derive s_D as a function of the separate group statistics. We begin by noting that the covariance between X and Y (g_{XY}) is defined by

$$g_{XY} = r_{XY} s_X s_Y = \frac{\Sigma XY}{N} - \frac{\Sigma X}{N}\frac{\Sigma Y}{N} \tag{15}$$

But in the Regression Projection Model

$$\frac{\Sigma XY}{N} = \frac{\Sigma X_t Y_t + \Sigma X_c Y_c}{N} \tag{16}$$

$$\frac{\Sigma X}{N} = \frac{\Sigma X_t + \Sigma Y_c}{N} \tag{17}$$

$$\frac{\Sigma Y}{N} = \frac{\Sigma Y_t + \Sigma Y_c}{N} \tag{18}$$

and

$$\frac{\Sigma X_t Y_t}{N} = P_t \frac{\Sigma X_t Y_t}{N_t} \tag{19}$$

$$\frac{\Sigma X_c Y_c}{N} = P_c \frac{\Sigma X_c Y_c}{N_c} \tag{20}$$

where P_t and P_c are the proportions of treatment and comparison students, respectively. Similarly

$$\frac{\Sigma X_t}{N} = P_t \frac{\Sigma X_t}{N_t} = P_t \bar{X}_t \tag{21}$$

$$\frac{\Sigma X_c}{N} = P_c \frac{\Sigma X_c}{N_c} = P_c \bar{X}_c \tag{22}$$

$$\frac{\Sigma Y_t}{N} = P_t \frac{\Sigma Y_t}{N_t} = P_t \bar{Y}_t \tag{23}$$

$$\frac{\Sigma Y_c}{N} = P_c \frac{\Sigma Y_c}{N_c} = P_c \bar{Y}_c \tag{24}$$

Substituting (19) through (24) in (16) through (18) and then the resulting equations in (15) we have

$$g_{XY} = \left\{ P_t \frac{\Sigma X_t Y_t}{N_t} + P_c \frac{\Sigma X_c Y_c}{N_c} \right\} - \left\{ (P_t \bar{X}_t + P_c \bar{X}_c)(P_t \bar{Y}_t + P_c \bar{Y}_c) \right\} \tag{25}$$

Next, we subtract the expression $(P_t \bar{X}_t \bar{Y}_t + P_c \bar{X}_c \bar{Y}_c)$ from the first brackets in (25) and add it to the second to get

$$g_{XY} = \left\{ P_t \left(\frac{\Sigma X_t Y_t}{N_t} - \bar{X}_t \bar{Y}_t \right) + P_c \left(\frac{\Sigma X_c Y_c}{N_c} - \bar{X}_c \bar{Y}_c \right) \right\} \tag{26}$$

$$+ \left\{ (P_t - P_t^2)\bar{X}_t \bar{Y}_t - P_t P_c \bar{X}_t \bar{Y}_c - P_t P_c \bar{X}_c \bar{Y}_t + (P_c - P_c^2)\bar{X}_c \bar{Y}_c \right\}$$

But we define

$$g_{XY_t} = \frac{\Sigma X_t Y_t}{N_t} - \bar{X}_t \bar{Y}_t \tag{27}$$

$$g_{XY_c} = \frac{\Sigma X_c Y_c}{N_c} - \bar{X}_c \bar{Y}_c \tag{28}$$

Also we have

$$(P_t - P_t^2) = P_t(1 - P_t) = P_t P_c \tag{29}$$

and similarly

$$(P_c - P_c^2) = P_c P_t \tag{30}$$

Using (27) and (28) in the first brackets of (26), and (29) and (30) in the second we have

$$g_{XY} = P_t g_{XY_t} + P_c g_{XY_c} + P_t P_c (\bar{X}_t - \bar{X}_c)(\bar{Y}_t - \bar{Y}_c) \tag{31}$$

Let

$$\bar{g}_{XY} = P_t g_{XY_t} + P_c g_{XY_c} \tag{32}$$

$$d_X = (\bar{X}_t - \bar{X}_c) \tag{33}$$

$$d_Y = (\bar{Y}_t - \bar{Y}_c) \tag{34}$$

Subtituting (32), (33), and (34) into (31)

$$g_{XY} = \bar{g}_{XY} + P_t P_c d_X d_Y \tag{35}$$

If Y = X, we have from (35)

$$s_X^2 = \bar{s}_X^2 + P_t P_c d_X^2 \tag{36}$$

Similarly, if X = Y

$$s_Y^2 = \bar{s}_Y^2 + P_t P_c d_Y^2 \tag{37}$$

Substituting (35), (36), and (37) into (14)

$$s_D^2 = b_c^2\left[\bar{s}_X^2 + P_t P_c d_X^2\right] + \left[\bar{s}_Y^2 + P_t P_c d_Y^2\right] - 2b_c\left[\bar{g}_{XY} + P_t P_c d_X d_Y\right] \tag{38}$$

Rearranging terms

$$s_D^2 = b_c^2 \bar{s}_X^2 + \bar{s}_Y^2 - 2b_c \bar{g}_{XY} + P_t P_c (d_Y - b_c d_X)^2 \tag{39}$$

Finally, it can be readily shown that

$$(d_Y - b_c d_X) = (\bar{Y}_t - \hat{\bar{Y}}_t) \tag{40}$$

and that

$$\bar{g}_{XY} = \bar{b}\bar{s}_X^2 \tag{41}$$

Substituting (41) and (42) in (40)

$$s_D^2 = b_c^2 \bar{s}_X^2 + \bar{s}_Y^2 - 2b_c \bar{b}\bar{s}_X^2 + P_t P_c (\bar{Y}_t - \bar{\bar{Y}}_t)^2 \qquad (42)$$

which is the form of the denominator in the equation for $\underline{t}$ on page 64.

C. The Generalized Multiple-regression Model

Where neither of the above models is indicated, it may be possible to apply a multiple regression model to the data, provided the evaluator can generate a useful null hypothesis. However, considerable caution and a thorough grasp of the technical issues involved should be considered prerequisites for any such effort. In particular, the widespread error of using regression models to statistically equate fundamentally dissimilar groups must be avoided. Campbell and Erlebacher (1970) have shown that, in terms of familiar "true score plus error score" models, conventional regression models systematically underadjust for the initial differences between such groups. More basically, it should be noted that the underlying "true score plus error score" construct is purely hypothetical and there is little evidence to suggest that it provides a useful basis for equating dissimilar groups. The behavior of one such group simply does not tell us much about the behavior of the other.

However, in special circumstances the Generalized Multiple-regression Model may prove to be applicable. In the simplest case, the first step in applying the model is to calculate a regression equation for the pretest-posttest distribution of the combined treatment/comparison group. The pretest score may be considered the "predictor" variable while the posttest score is the "criterion" variable. The variable of interest is the "residual variance;" that is, the posttest score variance which is not predicted by the pretest regression equation.

The second step is to add a "treatment" term as the second predictor in the regression equation and calculate the residual variance about the new regression line. In the simplest case, the treatment term is a dichotomous variable which would be given a value of "1" for each

student in the treatment group, and "0" for each student in the comparison group. There is, however, no reason why it could not be a continuous variable reflecting, for example, the hours of treatment exposure.

The last step is to test the significance of the difference between the residual variance computed from the first prediction equation, and the residual variance predicted from the second equation. The addition of the treatment variable in the second equation amounts to adding a constant to each treatment group score. Graphically, the result is to generate two parallel regression lines passing through the means of the treatment and comparison groups, respectively. The slope of these lines is the weighted mean of the independent regression lines for the two groups and will, in general, differ from the combined group regression line slope. The significance of the effect is determined by testing the difference between the residual variances from the two prediction equations.

The model is a "multiple" regression model in the sense that any number of predictors can be incorporated in the regression equation in addition to pretest and treatment variables (e.g., teacher ratings, SES, etc.). The model is "general" in the sense that a variety of effects can be examined singly, additively, and interactively. For example, by including a "treatment group" times "pretest scores" term it is possible to test whether treatment and comparison regression line slopes are significantly different. Finally, by including squared or other power terms, the shape of the regression line can be tested.

It will probably be recognized that the simple case described above is the Analysis of Covariance Model, a familiar special case of the Generalized Multiple-regression Model. The Y-axis distance between the two regression lines is the adjusted posttest difference. As indicated above, this difference will be a biased estimate if the groups are representative of distinct populations. A significant effect would provide a convincing (negative) answer to the question "Were the two groups of posttest scores drawn randomly from a single population?" However, such a conclusion

is trivial if it were known in advance that the groups were fundamentally different. Similarly, it is important in all applications of regression models to state the null hypothesis precisely, and to consider whether its rejection will be of any interest. Where there is any confusion concerning the assumptions of the null hypothesis or the implications of those assumptions, regression models cannot be recommended.

APPENDIX D

Hazards Associated with the Use of Percentiles and Grade-equivalent Scores

An important part of the development of commercial achievement tests is the collection of normative data from a large and usually nationally representative sample of students. These normative data permit the conversion of raw test scores into various types of "derived" scores (e.g., percentiles, stanines, grade equivalents) which provide useful frames of reference for interpretation. A percentile score, for example, provides an index of an individual pupil's status with respect to his age or grade-level peers. A grade-equivalent score is intended to equate an individual's raw score with the national average level of performance at some grade level.

Since all of these derived scores are based on national averages, it is essential that the sample of pupils tested be truly representative of the national population. It is also clear that the sample must be large enough so that random sampling errors are small and one can be confident that the statistics computed from the sample are very close to those which would have been obtained had the entire population been tested.

The importance of these sampling considerations is well known and amply documented (e.g., Angoff, 1971). Unfortunately, even if good normative data are collected by a test publisher there is no guarantee that the data will not be misused, misinterpreted, or both. In fact, the conventions adopted by test publishers in manipulating and reporting their normative data seem likely to enhance the probability of making various types of errors. It is these errors which are addressed here rather than the sampling considerations referred to above.

The normative data for many widely used commercial tests are collected during one short interval of the school year, usually either fall (e.g., Iowa Tests of Basic Skills, 1968 ed.), mid-year (e.g., California Achievement Test, 1970 ed.), or spring (e.g., SRA Achievement Series,

1971 ed.). While a few tests have empirical normative data points both fall and spring (e.g., Gates-MacGinitie Reading Tests, 1964 ed.; Stanford Achievement Tests, 1973 ed.), it is a common practice to generate derived scores through interpolation and extrapolation processes for times where no empirical data were collected.

If a test publisher were to collect normative data from nationally representative samples of children at all grade levels in the seventh month of the school year, it would be possible to construct tables for the seventh month of each grade level which enabled raw scores to be converted to their percentile equivalents. The raw score at the median of each grade-level distribution could also be appropriately converted to a grade-equivalent score. The median raw score of the first graders would thus correspond to a grade equivalent of 1.7, the median score of the second graders would correspond to a grade equivalent of 2.7, and so on. Both the percentiles and the grade-equivalent scores determined in this manner could be called empirical derived scores.

Clearly, if children are tested at the same time in the school year as the normative data were collected, it is possible to determine their percentile status with respect to the national sample. However, when children are tested at times which deviate from the empirical normative data points it is no longer possible to interpret percentile conversions meaningfully. It cannot be determined, for example, whether a child in the second month of second grade who scores at the fortieth percentile of children in the seventh month of second grade is above or below average with respect to his grade-level peers. Similarly, it is not possible to determine a grade equivalent for any raw score which does not correspond to the empirically determined median for grades 1.7, 2.7, etc. -- except by resorting to interpolation.

It is a relatively simple matter to generate additional grade-equivalent scores and percentile distributions by interpolating between empirical data points or by extrapolating beyond them. The assumptions underlying such projected derived scores, unfortunately, are tenuous at best and may be significantly in error. Before discussing projected scores,

however, it is useful to point out that even more serious errors can result from the failure to interpolate or extrapolate. The problem here is peculiar to percentile scores.

Most test publishers provide percentile norms for both the beginning and the end of the school year. Many also provide mid-year norms. It is either inferred or made explicit that the fall norms are "good" for September, October, and November; that the mid-year norms are good for December, January, and February; and that the spring norms should be used for testing dates in March, April, May, and possibly even June. The tables which present such norms enable one to convert test scores to percentiles or, conversely, to determine the test score which would, presumably, be obtained by children at any particular percentile position with respect to their grade-level peers.

Figure 3 was constructed from the norms tables provided by the Iowa Tests of Basic Skills, Form 5, Level 12. The solid line in Figure 3 shows the number of items which the test publisher says will be answered correctly by the median sixth-grade child at various times during the school year. It implies that all cognitive growth which takes place during sixth grade occurs overnight on November 30th and February 28th. The hypothesis that growth occurs in this manner is certainly untenable.

A more believable expectation for the cognitive growth of average sixth graders is shown by the broken curve which crosses the line representing the test publisher's "fiftieth percentile child" at mid-October, mid-January, and mid-April. If this line is taken to be a reasonable representation of the "real" median sixth grader, then comparison with the test publisher's "hypothetical" median sixth grader will show the real child below average at the beginning of each norming period and above average at the end of each period.

The amount of time-related distortion inherent in the norms is shown in Figure 4 where raw scores at the beginning and end of each normative period were taken from the broken line in Figure 3 and converted to percentiles using the test publisher's tables. In assessing

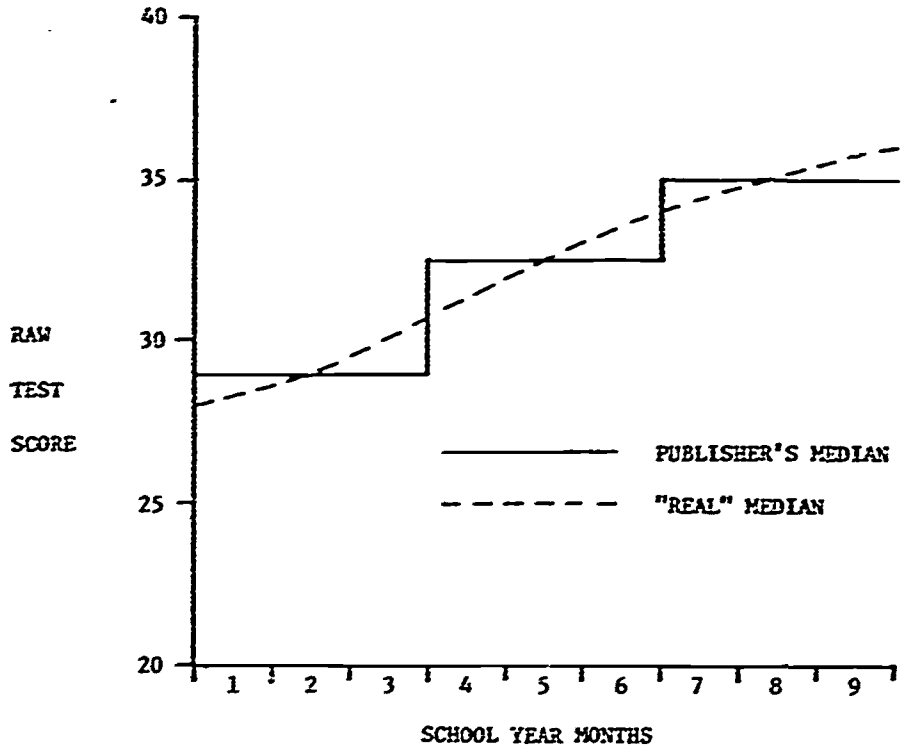


Figure 3. Cognitive growth shown by the test publisher's median versus a more realistic expectation

the progress of an individual student or the effect of a special instructional treatment, it is readily apparent that one would get results from pretesting early in a normative period and posttesting late which would differ dramatically from the results which would be obtained from the combination of late pretesting and early posttesting.

Where percentile norms are presented for the beginning, middle, and end of each school year, it seems highly likely that they are "correct" at some point in time within each of the three-month, nominal norm intervals. Those points in time, however, are unknown except in cases where empirical data have been collected. Where norms have been generated

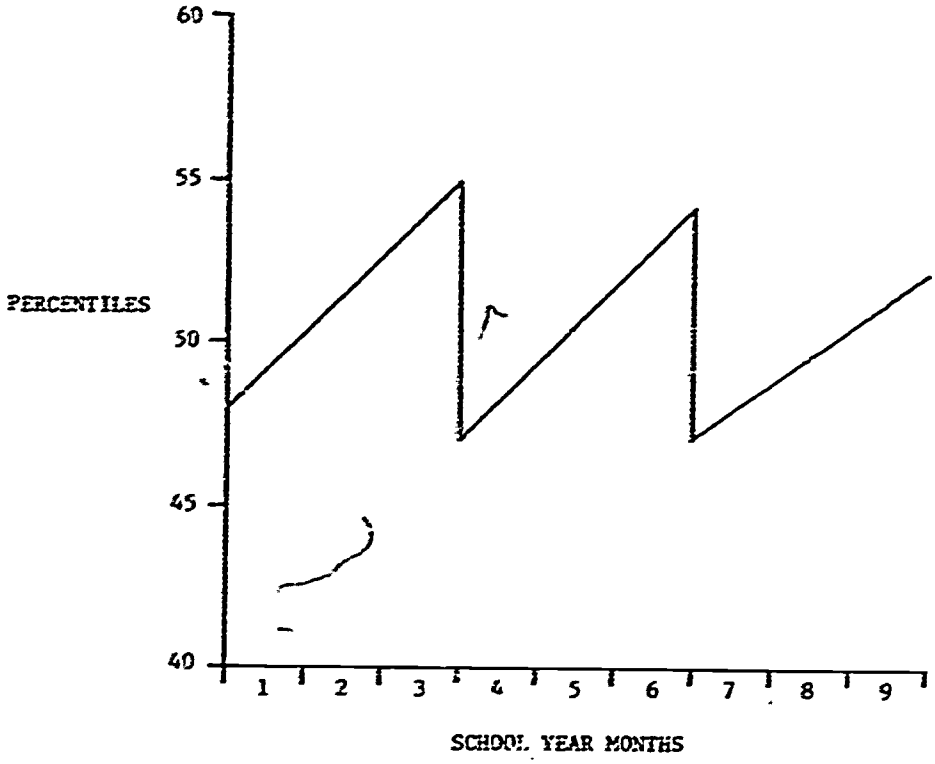


Figure 4. Publisher's percentiles corresponding to the "real" median in Figure 3 at the beginning and end of each norming period

through interpolation and extrapolation, it is probably safe to assume that the correct point is somewhere near the middle of the interval. However, any particular point which is chosen may be sufficiently in error to distort the findings of an evaluation study.

The same kind of problem exists with respect to grade-equivalent scores. These scores are usually derived as follows: (a) median raw score values are identified for each grade level at the month the test was normed (e.g., 1.7, 2.7, 3.7, etc.) and equated to these grade equivalents,

(b) the interval between medians is divided into ten equal parts, and (c) the intermediate grade-equivalent scores are equated with the nearest integral raw score value. The assumption which underlies this procedure, of course, is that the number of items answered correctly is a linear function of time over the nine months of the school year and that a third as much gain is made during each of the three summer months. This is essentially the same assumption which underlies projected percentile norms.

A number of studies have been undertaken to investigate the validity of the linear growth assumption, with perhaps the greatest amount of attention focused on the summer period where it appears most questionable. Findings have not always been consistent with respect to the direction of deviations between empirical and projected data points, but it is quite clear that such deviations are the rule rather than the exception. Wrightstone, Hogan, and Abbott in a recent publication (undated) of the Test Department, Harcourt Brace Jovanovich, Inc., concluded, "interpolated points may be considered as reasonably good estimates of the actual norms line if empirically determined points had been available for all times in the year. They are, however, almost certainly in error by some small amount in most cases and by a substantial amount in some cases [p. 8]."

Beggs and Hieronymus (1968) found different patterns of gains and losses with respect to the linear growth expectation on different subtests of the Iowa Tests of Basic Skills. They observed consistent and substantial summer losses in language and arithmetic areas but not in reading. Other deviations were noted but they were not consistent from grade to grade or even at different achievement levels within grade. They reported some evidence of accelerated growth from mid-January to mid-April in the language, work-study, and arithmetic areas.

Housley (1973), using the Stanford Achievement Test (1964 ed.), found that children showed neither gains nor losses from June of their third-grade year to the following September in either vocabulary or reading comprehension. Thomas (1975) reported similar findings from a study conducted in the San Jose, California school district, but Heyns (1975) reported reading achievement losses over the summer for blacks and low SES

students.

Some of the most interesting data can be found in the technical manuals of the test publishers -- particularly where tests have been normed twice during the school year and where both percentiles and grade-equivalent scores are presented. The issue of interest in these instances is that the fiftieth percentile child is not always at grade level! On the Metropolitan Achievement Tests (1970 ed.), for example, the median third grader is two months below grade level in reading at the end of the school year. Similarly, the median fourth grader is two months ahead of grade level in math at the end of the school year.

These anomalies result from a combination of two factors: (a) the conventions employed by test publishers in developing derived scores and (b) the fact that cognitive growth is not a linear function of time. It is standard practice, for example, to provide a single table converting raw scores to grade equivalents for each level of a test. To do so, of course, requires that the median child achieve a higher raw score at each successive point in time. A loss of raw score points over the summer would produce the interesting situation where a single score would correspond to three different grade equivalents. Figure 5 illustrates precisely this phenomenon.

The data plotted in Figure 5 are taken from the Norms Booklets, Form B, of the Stanford Achievement Test (Harcourt Brace Jovanovich, Inc., 1973). The data points connected by the solid lines represent the scaled scores in Mathematics Computation of the median child at grade levels 3.1, 3.8, 4.1, 4.8, 5.1, and 5.8 (raw scores had to be converted to scaled scores since the data were drawn from three levels of the test). The points connected by the broken line are scaled scores achieved by children scoring at grade level at the same points in time.

If the solid line in Figure 5 were used to convert scaled scores to grade equivalents, it can be seen that a score of 146 would convert to both 3.7 and 4.1. A scaled score of 147 would correspond to three different grade equivalents.

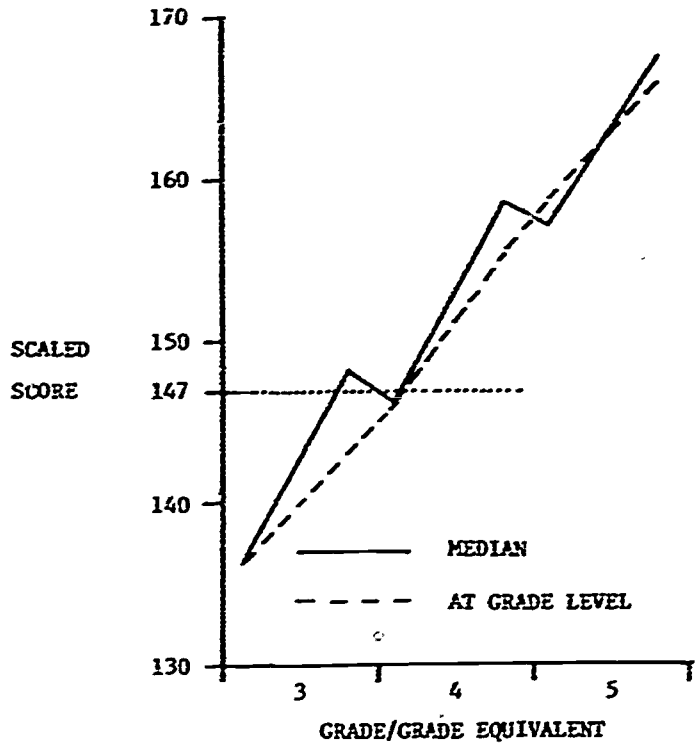


Figure 5. Comparison of the median score with the grade norm line

To avoid the confusion that might result from using a grade norms line such as the solid line in Figure 5, test publishers have adopted the convention of constructing a smoothed line to convert raw or scaled scores to grade equivalents. Such a smoothed line, of course, gives the mistaken impression that learning is a more orderly phenomenon than it really is and introduces distortions of sufficient magnitude to obscure whatever effects might result from any educational intervention. From the data reflected in Figure 5, for example, it can be shown that the third grader who scored exactly at the national average on both pretest and posttest would achieve grade-equivalent scores of 3.1 and 4.3 re-

spectively and would appear to have made a twelve-month gain in the seven-month period between the testings.

The example presented in Figure 5 is extreme, and other examples could be presented where the empirical data points correspond precisely with the projected points. Examples could also be presented where the distortion resulting from interpolation or extrapolation is in the opposite direction from that in the given example.

It should be clear from the above that projected grade-equivalent scores (and projected percentiles which reflect the same types of distortion) may deviate substantially from what they seem to be. Such scores will often not represent the median level of performance of children at the corresponding grade level. Furthermore, it can be shown that errors as large as several months are not uncommon.

Despite these problems, if it could be demonstrated that the errors in grade-equivalent scores were random with respect to the amount and direction of the distortion introduced, then it might still be possible to draw valid inferences regarding the effectiveness of educational programs under certain circumstances. Where such programs had been evaluated using several different test instruments at several different grade levels, for example, it might be safe to assume that the errors cancelled each other out and that mean grade-equivalent gains calculated across all pupils would be unbiased.

It is not possible, at the present time, to determine whether or not use of grade-equivalent scores to evaluate educational programs introduces systematic bias. To do so would require a demonstration that the gains made by median children (the national norm) were consistently non-linear over the ten-month school year. If the average gains per month were greater during that portion of the school year between fall and spring than between spring and fall, fall-to-spring grade-equivalent gains would be systematically inflated. Similarly, they would be systematically too low if the opposite pattern of gains prevailed.

The evidence cited above which found losses over the summer or gains

which were less than would be predicted under the linear growth assumption tend to support the hypothesis that grade-equivalent gains will be spuriously high from a fall pretest to a spring posttest. The findings, however, were not consistent with the possible exceptions of language and arithmetic. Certainly the research literature is not definitive on this issue with respect to reading.

Again, the normative data contained in the manuals accompanying tests with both fall and spring standardizations are relevant. They too, however, reveal an inconsistent pattern. The fiftieth percentile Total Reading score on the Metropolitan Achievement Tests (1970 ed.) is at grade level at the beginning of each grade and typically somewhat below grade level at the end of each grade. This pattern would result in grade-equivalent gain measures which systematically underestimated real cognitive growth.

Reading Comprehension scores on the Stanford Achievement Tests, Form A (1973 ed.), show exactly the opposite pattern. At every grade from first through eighth, the median fall score is below the grade norm line (grade level) and the median spring score is above it. Consequently, all fall-to-spring, grade-equivalent gains will be spuriously high.

A somewhat more consistent pattern can be observed in the test scores of children achieving below the national average. To illustrate this point, grade-equivalent scores on a variety of reading tests were drawn from the publishers' manuals for the 22nd percentile child. (This particular level was chosen because it is thought to be about the average for the ESEA Title 1 population.) Scores were collected for six instruments in all, at both fall and spring data points from grade 1.7 through 6.7. Grade-equivalent gains were computed for the fall-to-spring (school year) and spring-to-fall (summer) time intervals for each test. These gains were then divided by the number of school-year months in the interval to yield the average number of grade-equivalent months gained per school-year month.

Table 1 summarizes the gain data for the three tests which had empirical data points in both fall and spring (Gates-MacGinitie, 1964 ed.;

TABLE 1

Monthly Grade-equivalent Gains in Reading at the 22nd Percentile on Tests with Two Empirical Normative Data Points

Time Period	Gates	Metro	Stanford	Mean
First Grade				
Summer	.00	.50	.33	.28
Second Grade	1.00	.83	1.00	.94
Summer	.00	.50	-.33	.07
Third Grade	1.00	.33	1.00	.78
Summer	.33	.75	.33	.47
Fourth Grade	1.00	.83	.86	.90
Summer	1.00	.75	.00	.58
Fifth Grade	1.00	1.17	.93	1.03
Summer	1.00	.25	1.17	.81
Sixth Grade	.71	.83	.57	.70
Average Grade	.91	.80	.87	.87
Average Summer	.47	.55	.30	.44
Annual Expectation	.70	.70	.70	.70

Metropolitan Achievement Tests, 1970 ed.; and Stanford Achievement Tests, 1973 ed.) The scales represented are Total Reading for the MAT and SAT and Reading Comprehension for the Gates-MacGinitie (which does not provide Total Reading scores.) Averages calculated across grades and summers are presented for each test, and means calculated across tests are presented for each school year and each summer. The data labeled Annual Expectation are the mean monthly gains for each test over the entire period from the end of first grade to the end of sixth grade.

The most significant finding reflected in Table 1 is that, on the average, the monthly gain during the school year is almost exactly twice that which occurs over the summer. A child who maintains his status over the ten school-month period will average .87 months of grade-equivalent gain per school-year month from fall to spring and .44 months per month from spring to fall.

The same kind of analyses were carried out with three tests which have only one empirical data point per year, the California Achievement Test (1970 ed.), the Iowa Tests of Basic Skills (1971 ed.), and the SRA Achievement Tests (1971 ed.). The results of these analyses are presented in Table 2. It is interesting to note, in that table, that school-year gains are only about 30% higher than summer gains for these tests rather than 100% that was observed with those tests normed twice a year.

In attempting to interpret this difference, it is important to note that the basic raw-score-to-grade-equivalent conversion is probably not significantly more accurate for the double-normed tests than for those with only one empirical data point. The Metropolitan Achievement Tests interpolated grade-equivalent scores, in fact, were derived entirely from the fall data points in exactly the same manner as has generally been employed by test publishers when only one data point was available. The practice followed with the Stanford Tests was somewhat better but it, too, involved curve fitting and smoothing operations which clearly introduced some distortions.

Since the difference between the patterns of gains on the two sets of tests cannot be adequately explained in terms of the conversions tables,

TABLE 2

Monthly Grade-equivalent Gains in Reading
at the 22nd Percentile on Tests with
One Empirical Normative Data Point

	California	Iowa	SRA	Mean
Time Period				
First Grade				
Summer	1.25	.38	.25	.63
Second Grade	1.17	.95	1.33	1.15
Summer	.75	.15	.75	.55
Third Grade	.67	1.03	1.17	.96
Summer	.50	.63	.75	.63
Fourth Grade	.67	.92	.67	.75
Summer	1.00	.88	.50	.79
Fifth Grade	.83	1.00	.83	.89
Summer	.75	1.00	.75	.83
Sixth Grade	.50	.83	1.00	.78
Average Grade	.77	.95	1.00	.91
Average Summer	.85	.61	.60	.69
Annual Expectation	.80	.81	.82	.81

it has to result from the presence of empirical raw score distributions both fall and spring for one set of tests and not for the other. Where tests have only a fall or a spring empirical data point, the score distributions at the other period must b estimated by interpolation. The data in Tables 1 and 2 suggest rather strongly that the interpolation procedures used substantially overestimated gains from spring to fall and underestimated gains from fall to spring.

For 22nd percentile children who maintain their status with respect to their grade-level peers, Table 1 presents the grade-equivalent gains they can be expected to make on the Gates-MacGinitie, Metropolitan, and Stanford Achievement tests since the gains shown are all empirically determined. The gains shown in Table 2 on the other hand, are not empirically determined except over full-year periods. It is possible, however, to estimate how the average 22nd percentile child would score on the tests represented in Table 2 if he showed the same relative growth rates from fall to spring and spring to fall that were derived from the tests in Table 1. Such a child would have to gain 8 grade-equivalent months over the school year (Expectation from Table 2) while growing twice as fast from fall to spring as from spring to fall (Mean growth rates from Table 1).

If one assumes mid-October and mid-April testing dates, then the 22nd percentile child would, on the average, show a month-for-month gain from fall to spring (six months) and half-a-month-per-month gain from spring to fall (four months) when tested with the tests normed only once a year.

The conclusion that a 22nd percentile child would show month-for-month gains over the course of the school year while simply maintaining his status with respect to his grade-level peers seems intuitively nonsensical. It becomes shocking, however, when one considers that month-for-month growth is often taken to be the criterion of success in special compensatory education projects which supplement regular school experiences. To the extent that the analysis presented above is valid, month-for-month gains would be expected in the absence of any such special efforts!

The sum total of evidence presented in this appendix, while not entirely conclusive, suggests rather strongly that the obvious incongruity of

22nd percentile children making month-for-month gains does not result from the analytic step taken to arrive at that expectation but rather from the anomalies inherent in projected percentile distributions and grade-equivalent scores. Such scores appear to reflect both random and systematic errors of sufficient magnitude to invalidate any attempt to conduct a norm-referenced evaluation. If norm-referenced evaluations are to have any credibility whatsoever, they must be based entirely on empirical score distributions or projections of no more than a few weeks in either direction from such points.

Additional Problems with Grade-equivalent Scores

It might be argued that even though grade-equivalent scores systematically distort relationships between raw scores and empirically determined cognitive growth rates, the distortions are small enough so that they are more than counterbalanced by the advantages such scores possess with respect to simplicity and ease of understanding. The evidence presented above should be sufficient to dispel any illusions of this type as far as norm-referenced evaluations are concerned. The following discussion is intended to show that the apparent simplicity of grade-equivalent scores is entirely illusory and, furthermore, that they are scaled in such a way as to preclude their treatment with conventional statistical techniques.

The logical problems with grade-equivalent scores are well covered in many of the teachers' guides accompanying commercial tests. Specifically, a sixth grader who obtains a grade-equivalent score of four on a test is not really like a median fourth grader at all. Similarly, a second sixth grader who obtains a grade-equivalent score of eight is not like a median eighth grader. All that can be said is that these two sixth graders obtained the same scores that median fourth and eighth graders would have achieved on the sixth-grade test. Since their experiences, training, and intellectual growth rates have been very different from the students in higher or lower grades, it is not very meaningful to make implicit comparisons between them--particularly since these comparisons contain no information as to where the two children stand with respect to the achievement score distribution of their sixth-grade peers.

The interpretation of grade-equivalent scores is further complicated by the common misconception that being a year above or below grade level has the same meaning at different grade levels. Examination of the norms tables for any standardized achievement test clearly shows that this is not true. On the Metropolitan Achievement Tests, for example, a second grader who scores a year below grade level in Total Reading at the end of the school year is at the fourth percentile of the national distribution. A sixth-grade child scoring a year below grade level, however, is at the 38th percentile. The two points are separated by almost one-and-one-half standard deviations! It is also interesting to note that, according to the same norms tables, no children in first grade or the beginning of second grade are a year below grade level.

From a program evaluator's standpoint, the scaling problems are even more troublesome than the logical ones. The major difficulty is that the overall relation of achievement to school grade is not linear, as grade-equivalent scores would imply. The effect of this non-linear relation is illustrated schematically in Figure 6 for reading. No significance should be placed on the exact shape of the curve or the values in the figure. It is simply intended to suggest that the average student learns to read fairly well by the time he completes junior high school and thereafter makes relatively small gains in reading speed or comprehension (as distinguished from vocabulary).

The reading skill of the 50th percentile student in each grade, as measured on an achievement test, defines the grade-equivalent scores for the grade, so values on the reading-skill axis may be directly interpreted as the grade-equivalent values for each level of reading skill. It can easily be seen that, on this hypothetical curve, "half" the sixth-grade reading skill is represented not by a third-grade score, but by a second-grade score. Similarly, a fifth grader would be half way between third and ninth grade in terms of reading skill, while on a linear scale, the half-way point would be sixth grade.

While a curvilinear relationship between grade and skill level would be sufficient to invalidate most mathematical operations performed on

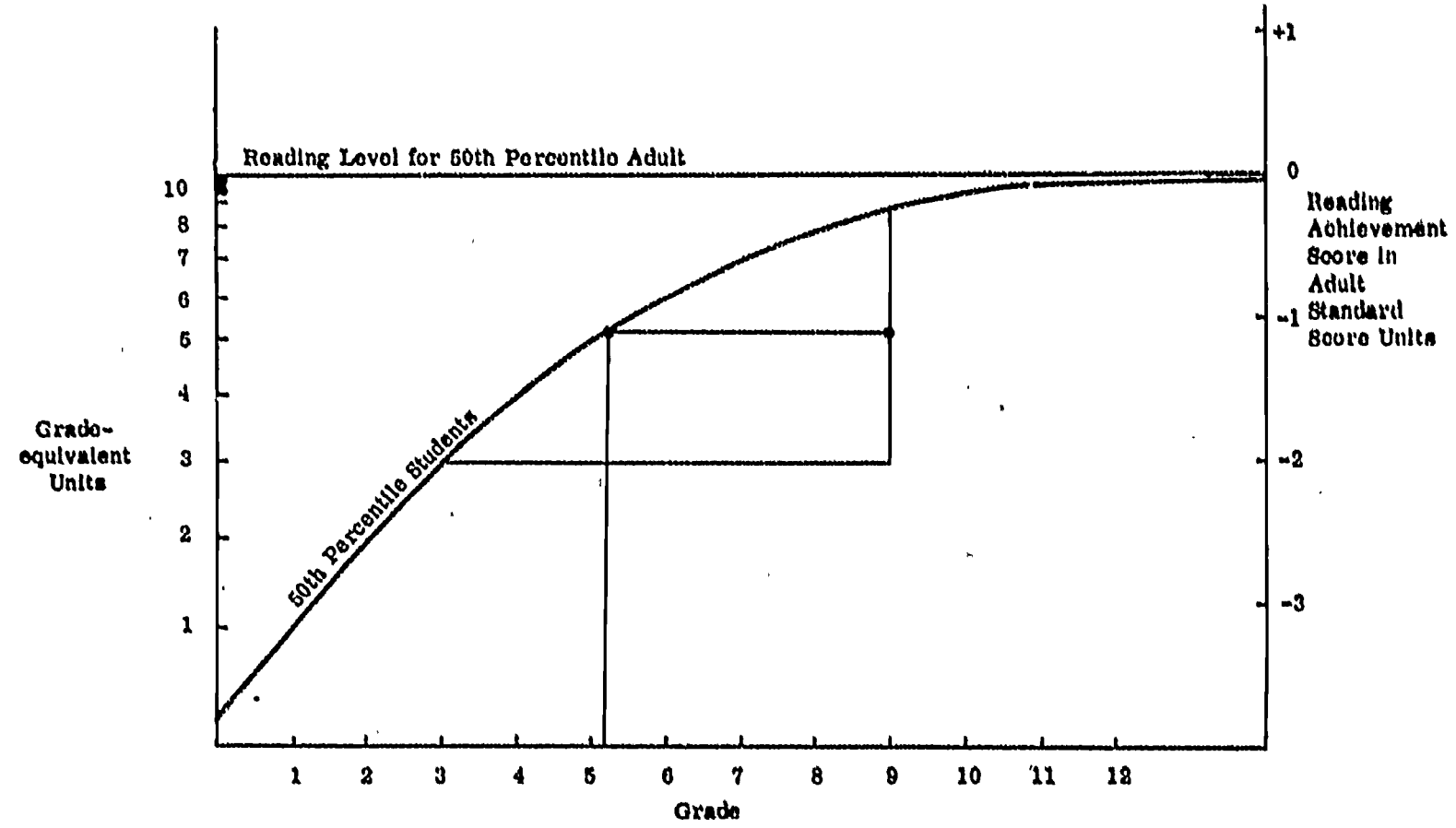


Figure 6. Hypothetical relationships between grade-equivalent score and reading skill.

TABLE 3

Mean Reading Comprehension Scores for Two Hypothetical Students on the Comprehensive Tests of Basic Skills (Form R)

	Raw Score	Scale Score	Grade Equivalent
Pretest – Grade 6.1			
Student A (16%ile)	15.00	396.0	3.70
Student B (84%ile)	34.00	573.0	9.20
Mean	24.50	484.5	6.45
Grade Equivalent	5.80	6.68	6.45
Error	-4.9%	-0.3%	+5.7%
Posttest – Grade 6.75			
Student A (16%ile)	17.00	415.0	4.10
Student B (84%ile)	35.50	592.5	9.75
Mean	26.25	503.0	4.10
Grade Equivalent	6.38	6.73	6.92
Error	-5.5%	-0.3%	+2.5%
Gain – Grade 6.1 to 6.75			
Student A (16%ile)	—	—	0.40
Student B (84%ile)	—	—	0.55
Mean	.58	.65	.47
Error	-10.8%	0.0%	-27.7%

grade-equivalent scores, there is some evidence that actual learning curves are considerably more irregular, and that curves for faster and slower learners are not necessarily the same shape as those for average learners. In general, averaging badly scaled grade-equivalent scores for students of different ability levels precludes any precise interpretation of group performance.

Table 3 presents an example of what can happen when scores on a non-equal interval scale are averaged. Two hypothetical students were chosen to present one standard deviation below the mean and one standard deviation above the mean, respectively, on the Comprehensive Test of Basic Skills (Form R) Reading Comprehension Scale. Normative data from grades 6.1 and 6.75 were arbitrarily selected. In this case, using the gain computed from standard scores as the "correct" gain, the mean grade-equivalent score underestimates the true gain by nearly two months. While the selected example is probably not typical of the effect, averaging a group of grade-equivalent scores will almost always yield a result which is substantially different from that which would be obtained by averaging the corresponding standard scores and then converting the mean standard score to a grade equivalent.

VI. REFERENCES

Airasian, P. W., & Madaus, G. F. Criterion referenced testing in the classroom. Measurement in Education, National Council on Measurement in Education, 1972, 3 (4), 1-8.

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

Beggs, D. L., & Hieronymus, A. N. Uniformity of growth in the basic skills throughout the school year and during the summer. Journal of Educational Measurement, 1968, 5 (2), 91-97.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Disadvantaged child. Vol. 3. Compensatory education: A national debate. New York: Brunner/Mazel, 1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.)

Davis, F. B. Criterion referenced measurement. ERIC Clearinghouse on Tests, Measurement, & Evaluation. Princeton, N.J.: Educational Testing Service, 1972, (Report 12).

Davis, F. B. Criterion referenced measurement. ERIC Clearinghouse on Tests, Measurement, & Evaluation. Princeton, N.J.: Educational Testing Service, 1973, (TM Report 17).

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), Psychological principles of system development. New York: Holt, Rinehart, & Winston, 1962.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971.

Guilford, J. P. Fundamental statistics in psychology and education. (4th ed.) New York: McGraw-Hill, 1965.

Harcourt Brace Jovanovich, Inc. Stanford Achievement Test, Manual Part II, Norms booklet, Form B. New York: 1973.

Heyns, B. Exposure and the effects of schooling. Berkeley, California: University of California. Technical Report under NIE Grant No. 30713, 1975.

Horst, P. Psychological measurement and prediction. Belmont, California: Wadsworth, 1966.

Horst, P. Effect of treatment as a special case of generalized multiple regression. Eugene, Oregon: Oregon Research Institute, 1974 (ORI Technical Report Vol. 14, No. 2).

Jackson, R. Developing criterion referenced tests. ERIC Clearinghouse on Tests, Measurement, & Evaluation. Princeton, N.J.: Educational Testing Service, 1971 (TM Report 1).

Levine, R. S., & Angoff, W. H. The effects of practice and growth on scores on the Scholastic Aptitude Test. Princeton, N.J.: Educational Testing Service, February 1958 (R and DR No. 58-6/SR-58-6).

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

Housley, W. Testing the "summer learning loss" argument. Phi Delta Kappan, 1973, 54, 705.

Parsons, H. M. What happened at Hawthorne? Science, 1974, 193, 922-932.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. In W. J. Popham (Ed.), Criterion-referenced measurement. Englewood Cliffs, N. J.: Educational Technology Publishers, 1971.

Porter, A. C. The effects of using fallible variables in the analysis of covariance. Unpublished doctoral dissertation, University of Wisconsin, 1967. (University Microfilms, Ann Arbor, Michigan, 1968).

Saretsky, G. The OEO P. C. experiment and the John Henry effect. Phi Delta Kappan, 1972, 53, 579-581.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

Sween, J. A. The experimental regression design—An inquiry into feasibility of nonrandom treatment allocation. Unpublished doctoral dissertation, Northwestern University, 1971.

Thomas, M. A. Cognitive growth over the summer and effects of homes on schools. California, Rand Corporation, 1974 (WN-8825-NIE).

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

Winer, B. J. Statistical principles in experimental design. (2nd ed.) New York: McGraw-Hill, 1971.

Wrightstone, J. W., Hogan, T. P., & Abbott, M. M. Accountability in education and associated measurement problems. Test Service Notebook 33. New York: Harcourt Brace Jovanovich, Inc. (undated).