DOCUMENT RESUME

ED 096 344 95 TM 003 950

AUTHOR Tallmadge, G. Kasten; Horst, Donald P.
TITLE A Procedural Guide For Validating Achievement Gains in Educational Projects. RMC Report No. UR-240.
INSTITUTION RMC Research Corp., Los Altos, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Office of Planning, Budgeting, and Evaluation.
REPORT NO UR-240
PUB DATE May 74
CONTRACT OEC-0-73-6662
NOTE 89p.

EDRS PRICE MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS *Academic Achievement; Criterion Referenced Tests; *Educational Programs; *Evaluation; *Models; Norm Referenced Tests; *Program Effectiveness; Research Design; Standardized Tests; Statistical Analysis

ABSTRACT

The orientation of this report is that of identifying educational projects which can be considered truly exemplary. The bulk of the report consists of a 23-step procedure for validating the effectiveness of educational programs 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. The report is concerned with deficiencies and hazards of various designs with emphasis on the weaker ones which, as it happens, are also the most feasible in real-world settings, the least costly, and the most commonly used. The appendixes contain project selection criteria worksheets, information regarding norm-referenced versus criterion-referenced tests, estimation of treatment effect from the performance of an initially superior comparison group, effects of noncomparable testing dates on experimental group versus norm group comparisons, and problems using grade-equivalent scores in evaluating educational gains. (Author/RC)

ED 096344

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: CA
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

RMC Report
UR-240

A PROCEDURAL GUIDE FOR VALIDATING ACHIEVEMENT GAINS IN EDUCATIONAL PROJECTS

G. Kasten Tallmadge
Donald P. Horst

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

May 1974

Prepared for
U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Office of Education/Office of Planning, Budgeting, & Evaluation

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

RMC Research Corporation
Los Altos, California

IM 003 950

ACKNOWLEDGEMENTS

The present version of this report is hopefully clearer, more accurate, and more insightful than the original draft for which the authors bore sole responsibility. Some of the changes were suggested by another six months of looking at real-world evaluation studies, and worrying about the many problems they failed to deal with in an adequate fashion. Of at least equal importance, however, were the many helpful comments and suggestions received from those who reviewed the draft report. We are particularly grateful to Edward B. Glassman, the U.S.O.E. Project Officer, both for his own thoughtful remarks and for those he solicited from his colleagues in the Office of Planning, Budgeting, and Evaluation, and elsewhere in U.S.O.E. We are indebted to Paul Horst for the very great assistance he provided with Appendix C of the report and for his many other comments and suggestions. Of the others who helped, Diane Jones is perhaps most deserving of special thanks for her patience, good humor, and editorial assistance in preparing, revising, and re-revising the manuscript.

G.K.T.

D.P.H.

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 clearly 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 intracacies of achievement test scores and normative data. Facts quickly came to light as this exploration proceeded which appeared to undermine the validity of inferences drawn from nearly all locally-conducted evaluations. The problems were so fundamental that the authors could not believe they were the first to discover them --

yet they were able to find nothing 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. The most recent 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 23-step procedure for validating the effectiveness of educational programs 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.

Some inferences may be drawn regarding the relative usefulness of various designs, but the report is really concerned with deficiencies and hazards. It follows, then, that emphasis is placed on the weaker designs which, as it happens, are also the most feasible in real-world settings, the least costly, and the most commonly used.

One additional point should be mentioned here. The orientation of this report is that of identifying educational projects which can be considered truly exemplary. It is necessarily less concerned with the non-selection of possibly successful projects than with the selection of possibly unsuccessful ones. 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 required.

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 are considered "possible" candidates for validation and all criterion areas were 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, but 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 page was filled in for every candidate project and, when completed, 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 page, "Preliminary Screening Criteria" comprises a checklist which was used for all projects which met the prerequisites. A project which clearly failed to meet any of the criteria was rejected without evaluating the other criterion areas, and, where doubt existed, effort was focused on the questionable area to avoid expending possibly fruitless effort on the others. Projects which survived the initial screening were subjected to additional investigation in all areas. Page three was used to summarize information resulting from these additional investigative steps in the availability, cost, and replicability criterion areas. Page four was used to describe the tryout design in such a way as to provide a context for considering the evidence of effectiveness.

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 grouped into the three categories of measurement, experimental design, and statistics. Hazards exist in each of these areas which may completely invalidate any inferences he might draw about 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 related problems of educational measurement. Scoring, scaling, and norming considerations become particularly important in those designs which employ non-comparable comparison groups or no comparison group at all.

The extent and complexity of the experimental and measurement problems made it clear that a systematic procedure was sorely needed 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 23-step decision tree was developed. The decision tree was designed to insure examination of each of the 12 threats to valid inference discussed by Campbell and Stanley (1963) as they relate to specific evaluation designs. It also encompasses other important considerations such as the type of scores on which statistical operations are performed (raw, standard, scale, percentile, grade-equivalent), whether comparisons are made against control groups or are norm-referenced, and the bases on which experimental-control (or norm group) comparisons are made (posttest scores, gain scores, covariance analysis, 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. These criteria include (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". There are associated measurement problems, however, which represent scientific challenges of a non-trivial nature. Most educators, for example, will agree 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 question, "How much closer?" must be answered by the policy makers. Once this criterion has been agreed upon, however, the problem of how to measure the improvement must be resolved.

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. Unfortunately, these fundamentally sound concepts do not stand up in practice.

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 the most severely disadvantaged children. This systematic bias stems from the fact that increasing an achievement growth rate from 0.9 to 1.0 months-per-month is clearly easier 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.

A criterion of this type must be defined in terms of an equal-interval scale with some sort of anchor point. Normalized standard scores referenced to a national average appear to offer the most appropriate medium in which such a criterion can be cast. Using unstandardized and/or criterion-referenced tests requires that success be defined in some other manner, and there can then be no assurance of equitability over the entire range of initial disadvantagement.

These considerations led the authors to advocate a definition of

educational significance which was expressed in terms of standard score gains referenced to the national norm. A gain of one-third standard deviation was subsequently agreed upon as the criterion to be used for determining exemplary status. Under these conditions, 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 gain are project specific, and in other projects even the desirability of equitability across all levels of initial disadvantagement might be offset by other considerations. The 23-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 page 46 summarizes the 23-step decision tree in flow-diagram form. Each step is discussed separately on the pages preceeding Figure 1. (This page arrangement is intended to facilitate reference to the fold-out figure.)

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 interpolation 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 23 which comprise the entire decision tree. Pages 5, 6, and 7 of Appendix A are worksheets for summarizing design characteristics and evaluation decisions.

One other point 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 Appropriate temporal stability reliability estimates should be used. In general, this means test-retest, or alternate forms estimates rather than measures of internal consistency such as split-half. Unfortunately, test-retest or alternate form reliability information is often omitted from test publishers' manuals. Reliability coefficients are seldom available for disadvantaged or other special groups. A rough reliability estimate for an experimental group with a restricted range of test scores (e.g., bottom 25%) may be obtained from:

$$r_{xx(exp)} = 1 - \frac{\sigma^2_{norm}\left[1 - r_{xx(norm)}\right]}{\sigma^2_{exp}}$$

This formula is based on the assumption that the error variance for the experimental group is equal to the error variance for the total norm group. If the experimental group error variance is actually higher than the total

norm group error variance this estimate of test reliability will be too high (see J. C. Stanley, 1971, p. 362). Floor effects will further lower reliability for a group in the tail of a distribution and a judgment must be made as to the magnitude of these effects (see Step 2).

The primary validity concerns are (a) whether the tests are sensitive to any gains students may be making (judgment based on comparison of the test content with program content is required) and (b) whether the tests are generally sensitive to improved reading or arithmetic skills. Widely recognized standardized tests may be accepted unless there appear to be glaring problems. Special purpose tests must be examined closely, and a judgment must be made. Appendix B discusses considerations relevant to criterion-referenced tests.

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 program is assessed by means of norm-group comparisons.

Step 2

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

Yes	Proceed to Step 3
No	Estimate the size of the effect, record on the worksheet, and 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. However, students often leave items blank 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 experimental 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 does not arise.

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 experimental group.

Step 3

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

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 may sensitize pupils as to what they are expected to learn. This sensitization may interact differentially with the learning experiences available to experimental 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 posttest scores of the program participants and could thus lead to erroneous conclusions regarding program impact. A compounding of this effect would almost certainly occur if pretesting was 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.

Step 4

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

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 experimental group or the "John Henry" effect 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 dissipate with time, the researcher has no real recourse but to rely on his own experience and judgment in deciding whether experimental outcomes should be attributed to treatment effects or simply to knowledge of group membership.

Step 5

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

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. Pupils for whom complete data are not available are likely to be systematically different from the others (lower socioeconomic status, more mobile families, higher absenteeism rate, higher dropout rate, etc.). For this reason, care must be exercised not to generalize the findings to the total group which was pretested.

Where pretest and posttest scores are reported on groups which are not identical (i.e., some children have pretest scores only and others have just posttest scores), systematic biases may be present. Students who dropped out, for example, may have been the lowest 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. If differential turnover is observed between the experimental and the control groups, explanations should be sought out and their impact on the experimental findings should be carefully assessed.

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 distinction between comparison and norm groups is not entirely clear cut, it is assumed that the data available on norm groups are cross-sectional in nature and do not include scores on individuals while data from typical control-group designs are longitudinal records of individual students. The latter are amenable to covariance analysis, while the former are not.

If some kind of control group is not employed in the evaluation design, gains made by the treatment group must be evaluated through norm-referenced comparisons. Comparisons of this type are usually reported in terms of either grade-equivalent gains or some measure of movement with respect to the national norm such as mean percentile shift. Such norm-referenced comparisons are discussed in the branch of the decision tree which begins with Step 7. Control group designs are discussed 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 a 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, the difference between the observed mean and the population mean must first be expressed in standard deviation units. The difference is then multiplied by the test-retest or alternate-form (not split-half) reliability coefficient presented in the test manual. The product may than be "translated" back into the units of the observed mean score to yield the estimated mean true score.

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. They should never be used for assessing the impact of educational influences.

Where real (as opposed to constructed) norms are used, they should be thought of as representing data from a control group. While even the most naive evaluators would recognize the folly of testing the experimental 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 7.6 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 work toward development of such norms is currently in progress, none, unfortunately, is yet available.

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 <u>are</u> 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 progress assessment.

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 norm will tend to move up at each grade level with respect to the non-dropouts, thus producing an apparent negative effect on their cognitive growth rates.

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 For gain scores to be useful in norm-referenced comparisons, it must be possible to relate them to the available normative data. A treatment group gain of 29 raw score points, for example, is uninterpretable unless it is possible to determine how large a gain was made over the same time interval by members of the norm group who were at the same initial achievement level when the experiment began. In other words, pretest scores must also be available.

Grade-equivalent gain scores appear to be an exception to this general rule. It seems that simply expressing gains in terms of grade-equivalent months per month of project exposure automatically provides 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 Appendices D and E).

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 23
No Skip to Step 13

Comment If normalized standard scores are available for both pre- and posttests, the significance of treatment-related, pre- to posttest change can be assessed by means of the following formula:

$$t = \frac{(\bar{X}_2 - \bar{X}_1) - (\text{Expected Gain})}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2}{N\ (N-1)}}}$$

where $\bar{X}_1$ = mean pretest score
$\bar{X}_2$ = mean posttest score
σ_1 = pretest standard deviation
σ_2 = posttest standard deviation
r_{12} = correlation between pre- and posttest scores
N = number of cases

Using this formula assumes, of course, that normative data (for the raw to standard score conversions) are available for testing dates comparable to the pre- and posttest administration times.

The expected gain is derived from normative data and represents an estimate of the cognitive growth project children

would have experienced in the absence of any special instructional treatment. It is zero where the normalizing and standardizing of scores has been done separately for each normative data point, a practice adopted most notably by the Gates-MacGinitie Reading Tests. A more common approach is to use expanded standard or scale scores which comprise a single scale extending across all normative data points and levels of the tests. Under these circumstances, a gain from pre- to posttest is expected even in the absence of a treatment effect. This expected gain must be subtracted from the observed gain in order to test the significance of the treatment effect.

The following steps should be executed in order to calculate the expected gain:

1. Determine the percentile equivalent of the mean pretest standard score.
2. Find the standard score equivalent of this percentile in the norm table appropriate for the posttest.
3. Subtract the mean pretest standard score from the standard score equivalent determined in Step 2. This difference is the expected gain.

Some test manuals provide simplified procedures for determining the significance of a gain from pre- to posttest. These procedures should not be used, however, as they incorporate assumptions about the correlation between pre- and posttest scores which may not be applicable to the program 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 It would be trivial to test simple gains from pre- to posttest since some cognitive growth is to be expected even in the absence of any special instructional treatment. What *is* of interest is that portion of the total observed gain which can be attributed to the special treatment. To find this gain it is necessary to derive the gain which would be expected under non-treatment conditions and subtract it from the total observed gain. The derivation of expected gains normally requires the availability of pre- or posttest scores (either, of course, can be derived from the other in combination with the mean gain). The single exception to this general rule is the case where scores have been normalized and standardized separately at each data point. Under this condition, the expected gain is zero regardless of what the mean pretest score may be. A far more common practice is to develop a single equal-interval scale which encompasses all data points and all levels of the test. With the notable exception of the Gates-MacGinitie Reading Tests, this latter approach has been adopted by all of the most commonly used achievement tests. Gain scores, under these circumstances, cannot be meaningfully interpreted without information about pretest score status.

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 23
No	Reject norm-group comparisons as adequate evidence of project success

Comment If the mean gain, the standard deviation (or standard error) of this mean, and the mean pretest score are available, the statistical significance of the gain can be assessed using the following formula:

$$t = \frac{\text{(Observed Gain)} - \text{(Expected Gain)}}{\text{Standard Error of the Gain}}$$

where the standard error is defined as the standard deviation divided by the square root of the number of children.

The expected gain should be derived from the mean pretest score according to the following steps:

1. Determine the percentile equivalent of the mean pretest standard score.
2. Find the standard score equivalent of this percentile in the norm table appropriate for the posttest.
3. Subtract the mean pretest standard score from the standard score equivalent determined in Step 2. This difference is the expected gain.

Step 14

Question Were the children, either matched or unmatched, randomly assigned to the experimental and control 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 experiment, 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 is employed, it should be implemented as follows. The entire pool of eligible children should be 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 be selected at random for assignment to the treatment group. The remaining member of the pair would, of course, be assigned to the control 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 experimental 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	Skip to Step 19

Comment 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]." Where pre-existing, intact groups are used as experimental 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 20
Gain Scores	Proceed to Step 17

Comment — Two types of gain scores are frequently used in educational evaluation, "raw", and residual gain scores. Comparisons between treatment and control groups based on raw gain scores (posttest scores minus pretest scores) are identical to comparisons based on posttest scores where the between-group posttest difference has been adjusted by the full amount of the pretest difference. Except in the case where the correlation between pretest and posttest scores is perfect, this adjustment is excessive. A pretest-posttest correlation of less than one implies that the pretest scores reflect some variance not included in the posttest scores. This variance, which is typically called measurement error, may reflect a large number of extraneous influences, some of which are random, while others represent systematic differences between the groups. In either case, variance due to measurement error is not relevant to post-test scores and represents a portion of the pretest scores which should not be subtracted from them. Since high pretest scores have a greater absolute amount of measurement error than low pretest scores, the use of raw gain scores will produce a spurious negative correlation between pretest status and gains. In other words, the higher the pretest score, the lower the gain. Thus, where the experimental group has lower pretest scores, the use of gain scores will increase the probability that a non-significant

treatment effect will appear significant. If the experimental group is initially superior, valid inferences may be drawn about the treatment effect <u>if</u> the raw gain scores of the two groups are found to be significantly different.

Residual gain scores are not gain scores at all but are differences between observed posttest scores and posttest scores predicted from the regression of posttest on pretest scores for the experimental and control groups combined. Where the regression line for the combined group is a weighted average of the within-group regression lines, residual gain scores are equivalent to posttest scores adjusted for pretest differences through covariance analysis. This equivalence, however, does not hold except where the two groups have comparable pretest score distributions. In fact, where pretest scores are substantially different and posttest scores are equal, the slope of the combined-group regression line approaches zero and the residual gain technique obscures the effect of pretest differences completely. Since residual gain scores systematically under-correct for pretest differences, their use is always undesirable. Where analysis of residual gain scores indicates that an initially inferior treatment group has outperformed the comparison groups, the success of the treatment can be accepted. Where results under these circumstances are non-significant, or where the treatment group scored higher than the controls on the pretest, the results of the analysis should be regarded as inconclusive at best. There is a very real danger that a successful treatment will be rejected using this procedure and some other form of analysis should be undertaken if at all possible.

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 23
No	Skip to Step 21

Comment Wherever pretest differences between experimental and control groups have resulted from random assignment procedures, covariance analysis should be employed, if possible, 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.

Step 18

Question Were pretest scores collected?

Yes	Go back to Step 15
No	Proceed to Step 21

Comment If assignment of pupils to experimental 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.

Step 19

Question Is the control group superior to the experimental group on the balance of educationally relevant variables?

Yes	See Appendix C
No	Reject control group comparison as adequate evidence of project success

Comment "Educationally relevant variables" include, but are probably not limited to, pretest scores, socioeconomic status, age, sex, racial and ethnic composition, and school and community factors. Where there are significant differences between experimental and control groups on one or more of these variables, "true" experimental designs cannot be employed. The alternative quasi-experimental approaches which may be adopted all rest on sets of assumptions of varying degrees of plausibility. If it could be assumed, when dealing with non-comparable groups, that they would respond in a similar manner to the presence or absence of the variable under investigation, there would be no real problem. Because this assumption is untenable, however, it is generally necessary to make other assumptions about how their responses would differ. One such assumption which is relevant here and appears "safe" is that a group which is initially superior to another group in cognitive development will continue to grow at a rate equal to or greater than that of the initially inferior group, other things being equal. If, then, the initially inferior group outperforms the initially superior group after exposure to a special educational treatment, it is probably safe to conclude that the treatment was effective. On the other hand, if the treatment had been administered to the initially superior group, it would not be possible to reach

any conclusion by comparing its growth rate against that of the initially inferior group.

Other assumptions could be made which would permit the quantification of growth rates and thus enable comparisons to be made in both directions and with the appearance, at least, of greater precision. Assumptions of this type, unfortunately, tend to require massive doses of faith since there is little in the way of empirical data to support them. Until such data are assembled, the possibly arbitrary decision has been made to reject as inadequate any experimental-control group comparisons where the experimental group is initially superior to the control group.

Step 20

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

Yes	Skip to Step 23
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), analysis of covariance is the most powerful 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 not be regarded as a substitute for truly comparable 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 error 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 in large.

Step 21

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

Yes Skip to Step 23

No Proceed to Step 22

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 educational programs are likely to be badly skewed.

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 valid inferences from gain-score comparisons discussed in Step 16 should be carefully observed.

Step 22

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

Yes	Obtain data, compute appropriate statistics, and proceed to Step 23
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 21.

Step 23

Question Do analysis results favor the treatment group at the pre-selected 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 determining the statistical significance of a treatment effect requires careful consideration of each of the extraneous effects identified in proceeding through the decision tree and an attempt to estimate 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.

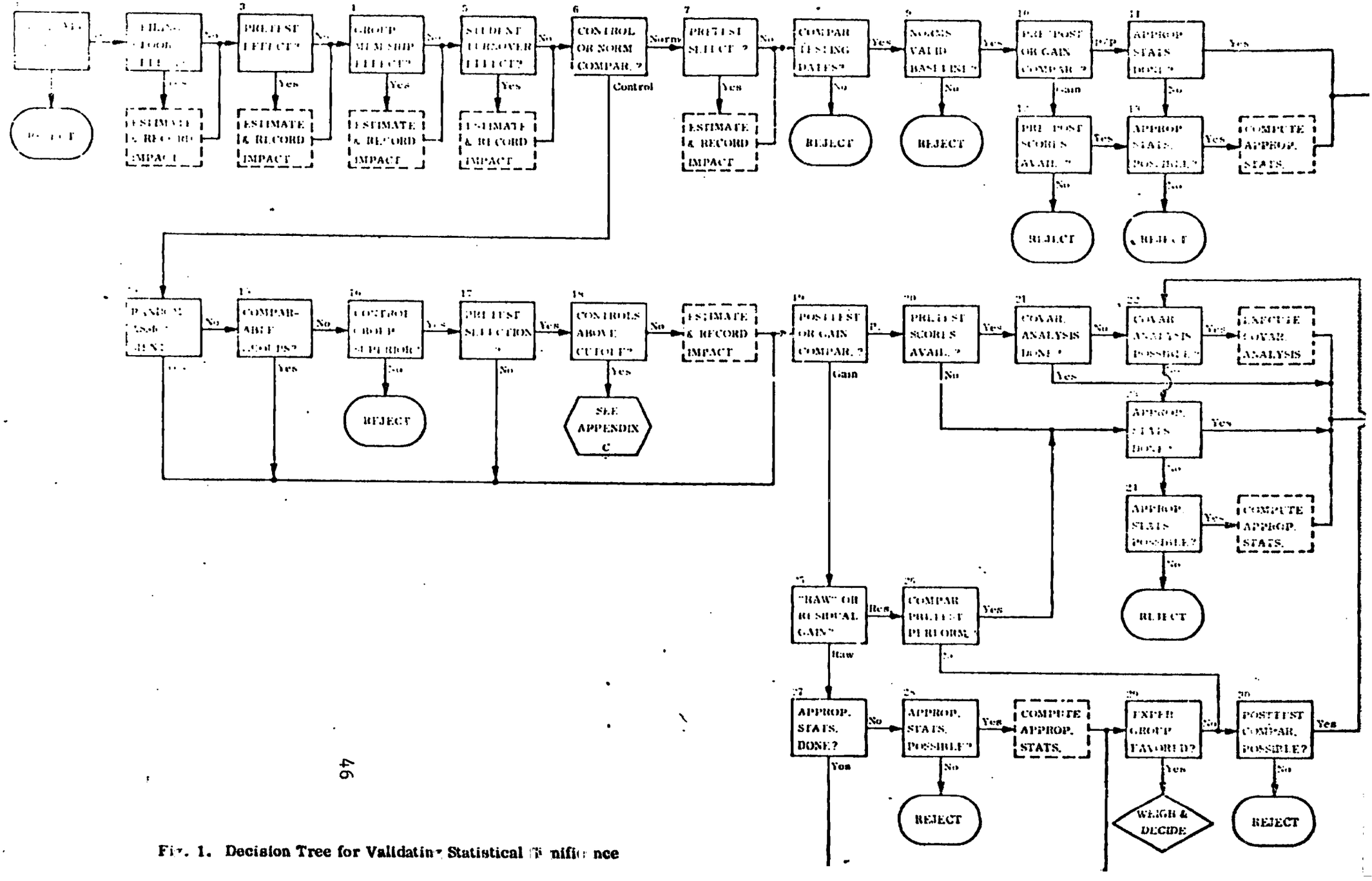
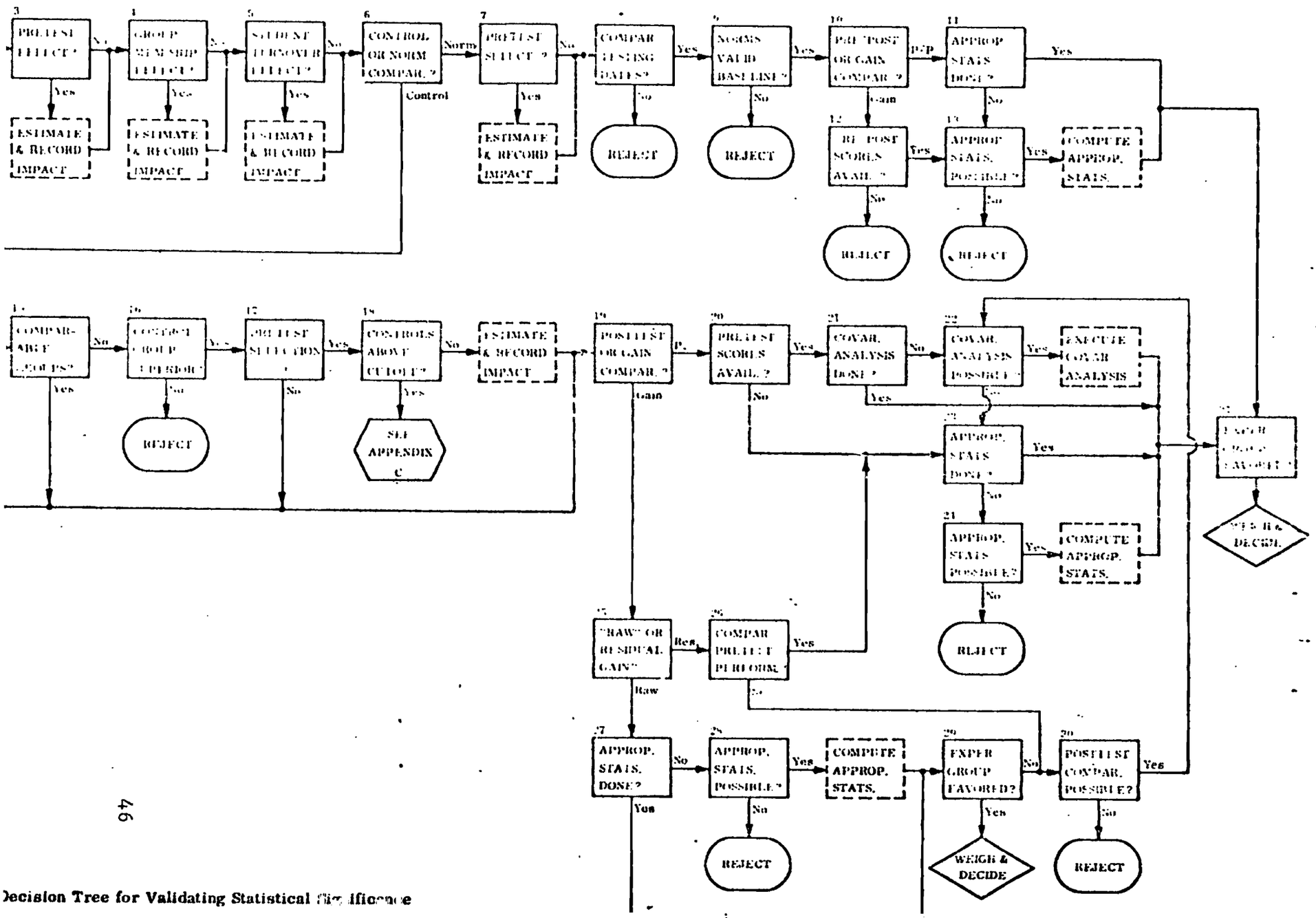


Fig. 1. Decision Tree for Validating Statistical Significance



Decision Tree for Validating Statistical Significance

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 subtests 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, $r \approx .80$) 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 educational programs 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

SUMMARY PAGE

PROJECT TITLE_____

Date	Initials	
		DESCRIPTION Approach Pull-out - Whole class
		PREREQUISITES Content Grades Tryout population Number of tryouts
		PACKAGING CRITERIA I. Availability II. Cost III. Replicability IV. Effectiveness Statistical Significance Educational Significance

PROJECT SELECTION CRITERIA WORKSHEET

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 $400 per pupil
- ☐ Initial investment less than $1000 per pupil
- ☐ (Alternatively) Per-pupil cost over a three year operational period including start-up costs should not exceed $735

REPLICABILITY

- ☐ Operating programs are provisionally considered replicable unless a major component clearly cannot be readily duplicated. Components include: materials, hardware, personnel, and environments.

EFFECTIVENESS

- ☐ Norm and/or control group with comparable test dates

PROJECT SELECTION CRITERIA WORKSHEET

NOTES

AVAILABILITY

COST

REPLICABILITY

PROJECT SELECTION CRITERIA WORKSHEET

NOTES

EFFECTIVENESS

Description of tryout design(s)

PROJECT SELECTION CRITERIA WORKSHEET

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. Experimental 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

ANALYSIS OF PROJECT EVALUATION

C. Norm-referenced (standardized) tests

Name	Exp/Cont	Pretest		Posttest		Data reported
		Date	N	Date	N	

D. Other measures (student, teacher, parent, other)

Criterion-referenced tests
Intermediate/Formative data
Opinion/Attitude data
Critical incidents
Classroom grades
Attendance/Discipline records
Other

PROJECT SELECTION CRITERIA WORKSHEET
ANALYSIS OF PROJECT EVALUATION

II. Evaluation of Effectiveness

A. Factors affecting statistical significance

1. Adequate tests

2. Ceiling/Floor effects

3. Pretest effect

4. Group membership effect

5. Student turnover

6. Experimental/Control analysis steps

7. Experimental/Norm analysis steps

B. Educational Significance

C. Other outcomes; unexpected outcomes

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 frequent 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.

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-ferenced." 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 is 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 the usual norm groups are not typical of disadvantaged students (see Step 9 of the decision tree) and the experimental groups are not tested at the same time of year as the norm groups (see Appendices D and E).

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 no reason to believe that tests developed according to "criterion-referenced" procedures provide better measures of project effectiveness in basic skill areas. Commercial tests clearly sample important aspects of reading and math achievement and are relatively efficient and reliable instruments. They also provide normative data that permit comparisons among projects. However, "criterion-referenced" or other special-purpose tests may be used to assess project effectiveness if enough is known about their properties to justify estimating the significance of gains. One requirement, of course, if that both the statistical and educational significance of observed gains must be assessed against the gains which would be expected under non-treatment conditions. In the absence of normative data, the computation of expected gains clearly necessitates the use of a control group evaluation model.

APPENDIX C

Estimation of Treatment Effects from the Performance of an Initially Superior Comparison Group

In many educational evaluations the only available comparison group is one which is initially superior to the treatment group. Using conventional statistical tests, the treatment group must surpass the comparison group by a statistically significant amount on the posttest before it can be inferred that the treatment was effective. This requirement is unduly restrictive and could result in the rejection of many successful projects, especially where pretest differences were large. There are, however, several quasi-experimental regression models which are applicable at least in certain instances and which may permit reasonably convincing conclusions to be drawn. Where the required data are available and the effort appears warranted, application of one of these models may be indicated. Three such models are discussed below.

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 below a fixed cutoff score to the treatment condition and all members above the cutoff to the comparison group.[1] Separate regression lines are then computed for each group and the

---

1. Step 19 of the decision tree requires that a non-comparable control group be initially superior to the treatment group. This restriction is not strictly relevant to the regression-discontinuity model which could be applied equally well to the evaluation of special programs for gifted students where the comparison group was initially inferior.

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 pretest scores, or even of scores reflecting both pretest performance and a teacher rating. Second, the model is not sensitive to changes in regression line slopes unless these changes are accompanied by a discontinuity of the regresssion lines. This requirement represents at least 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 effecting a flattening of the treatment-group regression line slope not accompanied by 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.

The remaining two models offer wider applicability at the expense of entailing more tenuous assumptions. The simpler of the two uses a regression line calculated from the control 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 control components about a particular pretest cutoff score. The model owes its origin to the technique of Karl Pearson for estimating total group test validity when criterion measures are available only for those who scored above some selected cutoff point. The

estimation technique is exactly analagous to that dealt with here. Selection (pretest) scores are available for an entire group but 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 which scored above the cutoff.

Horst (1966), Chapter 26, provides a discussion of the issues and presents formulas for generating unbiased estimates of the mean, standard deviation, and pretest-posttest correlation for the total group based on the following statistics: (a) the mean pretest score of the total group, (b) the mean pretest score of the restricted (control) subgroup, (c) the standard deviation of pretest scores for the total group, (d) the standard deviation of pretest scores for the restricted (control) subgroup, (e) the standard deviation of posttest scores for the restricted (control) subgroup, and (f), the correlation between pre- and posttest scores for the restricted (control) subgroup.

Given the estimated total-group statistics, a total-group regression equation can be developed. This equation can then be used to compute estimated posttest scores from pretest scores of the treatment group. These scores will provide an unbiased estimate of how the treatment group would have done had they not received the treatment--and there need be no concern with regression effects, since these influences are automatically considered by the procedure.

In practice, of course, there is no need to calculate the posttest mean or standard deviation or the pretest-posttest correlation for the total group in order to get a total-group regression equation. The estimated regression equation for the total group is identical to the regression equation for the restricted (control) group. Thus, one needs only to calculate the regression equation for the control group and use it to obtain estimated posttest scores.

The basic equation for predicting treatment group posttest scores is:

$$\hat{Y}_T = X_T b_C + k_C$$

where b_C is the slope of the control-group regression line and k_C is the Y-axis intercept.

If the mean pretest score of the treatment group is substituted for X_T in the above equation, $\hat{Y}_T$ will be the estimated mean posttest score. The difference between the observed and estimated posttest scores can then be tested using the following formula:

$$F_{1,N-3} = \frac{p_T (\bar{Y}_{obs} - \bar{Y}_{est})^2 (N-3)}{\bar{\sigma}^2_y - 2b_C\bar{b}\bar{\sigma}^2_x - b^2_C\bar{\sigma}^2_x + p_T p_C (\bar{Y}_{obs} - \bar{Y}_{est})^2}$$

where p_T = proportion of pupils in the treatment group

p_C = proportion of pupils in the control group

N = number of pupils in the combined group

$\bar{\sigma}^2_y$ = weighted average of the treatment- and control-group posttest variances

$\bar{\sigma}^2_x$ = weighted average of the treatment- and control-group pretest variances

b_C = slope of the control-group regression line

$\bar{b}$ = weighted average of the slopes of the treatment- and control-group regressi lines

Less powerful but computationally simpler non-parametric tests are also appropriate (e.g., Kolmogorov-Smirnov).

The third approach employs a generalized multiple-regression model derived from a statistical test in Winer (1971, p. 141). Of the three models presented in this appendix, it is the only one which does not require the development of experimental and control groups by dichotomizing a single original group at a specific pretest cutoff score. A complete mathematical development of the model with appropriate tests of

significance is provided by Horst (1974).

In the simplest case, the first step in applying the generalized multiple-regression 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 control 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 control 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 Y-axis difference between the lines can be interpreted as the treatment effect. This treatment effect is algebraically equivalent to the effect calculated in analysis of covariance (see Winer, 1971, p. 768). 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 control regression line slopes are significantly different. Finally, by including squared or other power terms, the shape of the regression line can be tested.

While the two latter models have somewhat broader applicability than the regression-discontinuity model, a corresponding increase in caution is indicated in their use and interpretation. In their simplest forms, both models essentially test the significance of deviations from a single straight regression line. The basic assumption is that the regression line would have been straight had there been no treatment effect. The plausibility of this assumption, however, should be subjected to careful scrutiny before a significant departure from a straight line is construed as evidence for a treatment effect. The shape of the pretest-posttest regression line is a function of the test properties, scaling technique, and the similarity of the tested group to the norm group. Unfortunately, the values of these variables are not readily available.

Given that the line is straight, it is also necessary to assume that the slope and intercept of the "no treatment" regression line vary from group to group (and must therefore be estimated from the non-comparable controls). This assumption also lacks empirical support. One may argue that if the slope varies from group to group, then it may also vary from subgroup to subgroup within a given group, resulting in a curved total-group regression line even without a treatment effect. This argument must be answered independently for each test instrument and requires the examination of regression lines for groups in which all students receive homogeneous treatment. In short, while these two models provide powerful tools for identifying a variety of potential treatment effects, their use in selecting among alternative hypotheses requires both a thorough understanding of the assumptions of the models and reasonable assurance that the assumptions are met.

APPENDIX D

Effects of Non-comparable Testing Dates on Experimental Group versus Norm Group Comparisons

An important part of the development of commercial achievement tests is the collection of normative data from a large sample of students. The normative data permit the transformation of raw scores into percentile scores, standard scores, or grade-equivalent scores which provide useful information about the meaning of individual raw scores in relation to the particular norm group in question. The importance of having a relevant norm group is discussed in Section IV of this report. The importance of having comparable test dates for experimental and norm groups is also referenced there but is discussed here in greater detail.

There is convincing evidence that learning, as reflected in achievement test scores, is typically not uniform over the calendar year (Beggs & Hieronymus, 1968). It is not possible to generalize as to the nature or causes of the non-uniformity, but one widely recognized factor that appears to operate in certain situations is the effect of "forgetting" over the summer months. This effect is illustrated by the hypothetical "observed score" line in Figure D1.

The normative data for many widely used commercial tests are collected during one short interval of the school year, typically February or March (e.g., California Achievement Test, Comprehensive Test of Basic Skills, SRA Achievement Series, Stanford Achievement Tests). In order to estimate appropriate scores for fall and spring, the single data points from successive years are simply connected with a smooth curve as illustrated by the broken line in Figure D1. It is obvious that, for the hypothetical data in the figure, this procedure systematically overestimates the expected fall score and underestimates the expected spring score. It should be clear that if the estimated fall and spring scores were used as the comparison standards for a special instructional program,

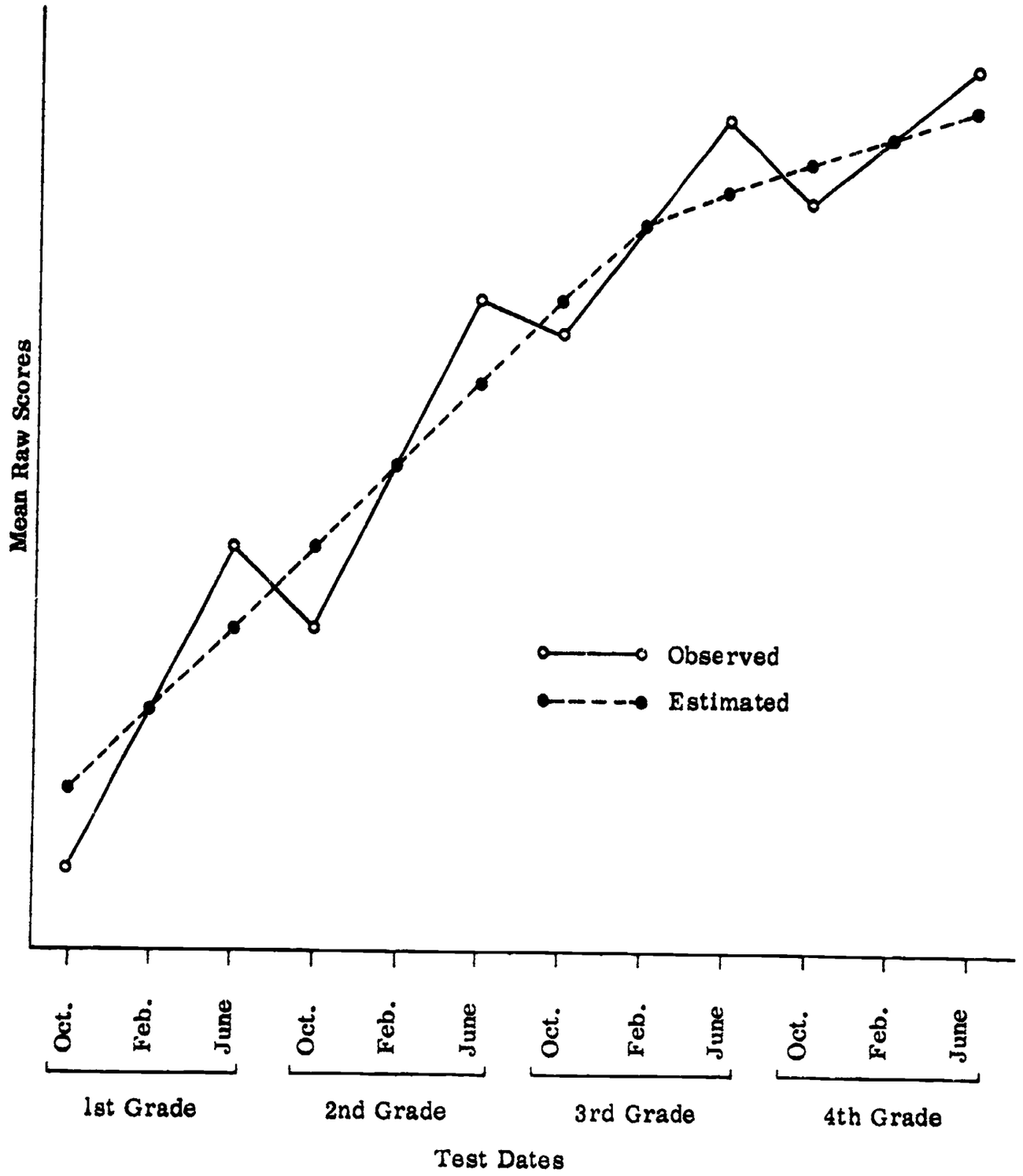


Fig. D1. Hypothetical achievement test scores.

the program might appear to give unusually good results when actually the improvement was exactly the same as that achieved by an "average" group of students. It can be seen that observed norm-group mean scores in October are far below the estimated scores. This means that an experimental class scoring exactly at the norm-group fall mean would appear to be doing very poorly when compared to the estimated fall norm-group mean. In the spring, assuming that they continue to do exactly as well as the norm group, the experimental class would score well above the estimated spring score. In fact, if the estimated fall and spring scores of Figure D1 were used to assess the progress of a typical norm-group class during a given school year, one would get the erroneous impression that a very poor class had been transformed into a very good class.

All types of scores which are estimated by interpolation between data points are likely to introduce systematic errors into educational evaluations. These include, in general, standard scores, percentile scores, stanines, and grade-equivalent scores. Grade-equivalent scores are characterized by additional problems which are discussed in detail in Appendix E. Even expanded standard or scale scores may be somewhat distorted by curve-fitting procedures required to achieve articulation between levels of a test.

It must be emphasized that the data points in Figure D1 are purely hypothetical and that different, conceivably even opposite effects might be found with specific tests or norm groups if the data were available. However, in the few tests which do report normative data from two points during the year (e.g., Gates-MacGinitie Reading Tests, and Metropolitan Achievement Tests) the effect illustrated in Figure D1 does appear to be present (see Appendix E, Figure E5). The implication of these data is that tests which provide normative data for only one point in the year should not be used for norm-referenced evaluation of fall-to-spring gains, and that, in general, it is not advisable to extrapolate or interpolate very far from observed normative data.

APPENDIX E

Problems With Using Grade-equivalent Scores in Evaluating Educational Gains

Evaluation reports for experimental educational projects frequently present results in terms of grade-equivalent scores or grade-equivalent gains. The apparent simplicity and ease of interpretation of grade-equivalent scores has probably been responsible for their widespread adoption. Unfortunately, however, this apparent simplicity is entirely illusory, and there is ample evidence to contraindicate the use of grade-equivalent scores or grade-equivalent gains for any purpose whatsoever in educational evaluation.

The problems with grade-equivalent scores can be divided into logical and scaling considerations. The logical considerations 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 as median fourth and eighth graders reading sixth-grade material. 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.

From a program evaluator's standpoint, the scaling problems are even more troublesome than the logical ones. There are two primary considerations: first, the overall relation of "reading skill" to "school grade" is not linear as grade-equivalent scores would imply. This makes the computation of mean grade-equivalent scores inappropriate. Second, the

relation of "reading skill" to "school grade" is not well behaved (i.e., not smooth) over short sections of the curve. The typically jagged norm-group data curve is difficult to work with so test developers usually do some "smoothing." This smoothing introduces systematic inaccuracies when grade-equivalent scores are used in a project evaluation. The effects of these two kinds of problems, as well as several others, are illustrated in the following discussion by hypothetical data, and by actual curves from published reading comprehension scales.

The effect of the non-linear relation between reading skill and school grade is illustrated schematically in Figures E1 and E2. Figure E1 illustrates the commonly used format for graphically representing student progress in terms of grade-equivalent scores. The apparent simplicity of this format obscures important fundamental information about the acquisition of skills such as reading which are typically learned up to a certain level, and then maintained at that level throughout adulthood.

The format of Figure E2 is probably more appropriate for representing reading achievement. 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

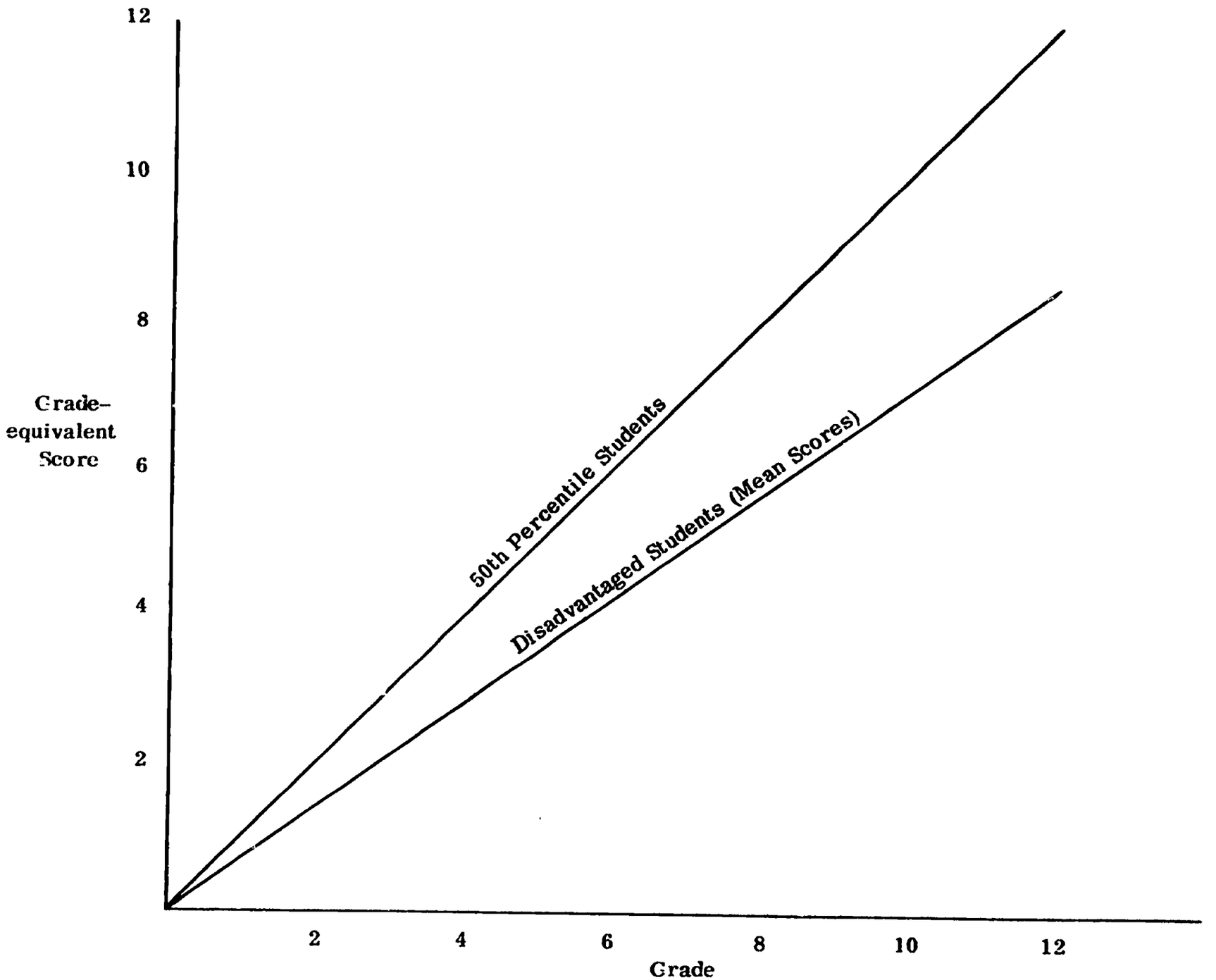


Fig E1 ... othetical curve illustrating typical for ... for ... groups ...

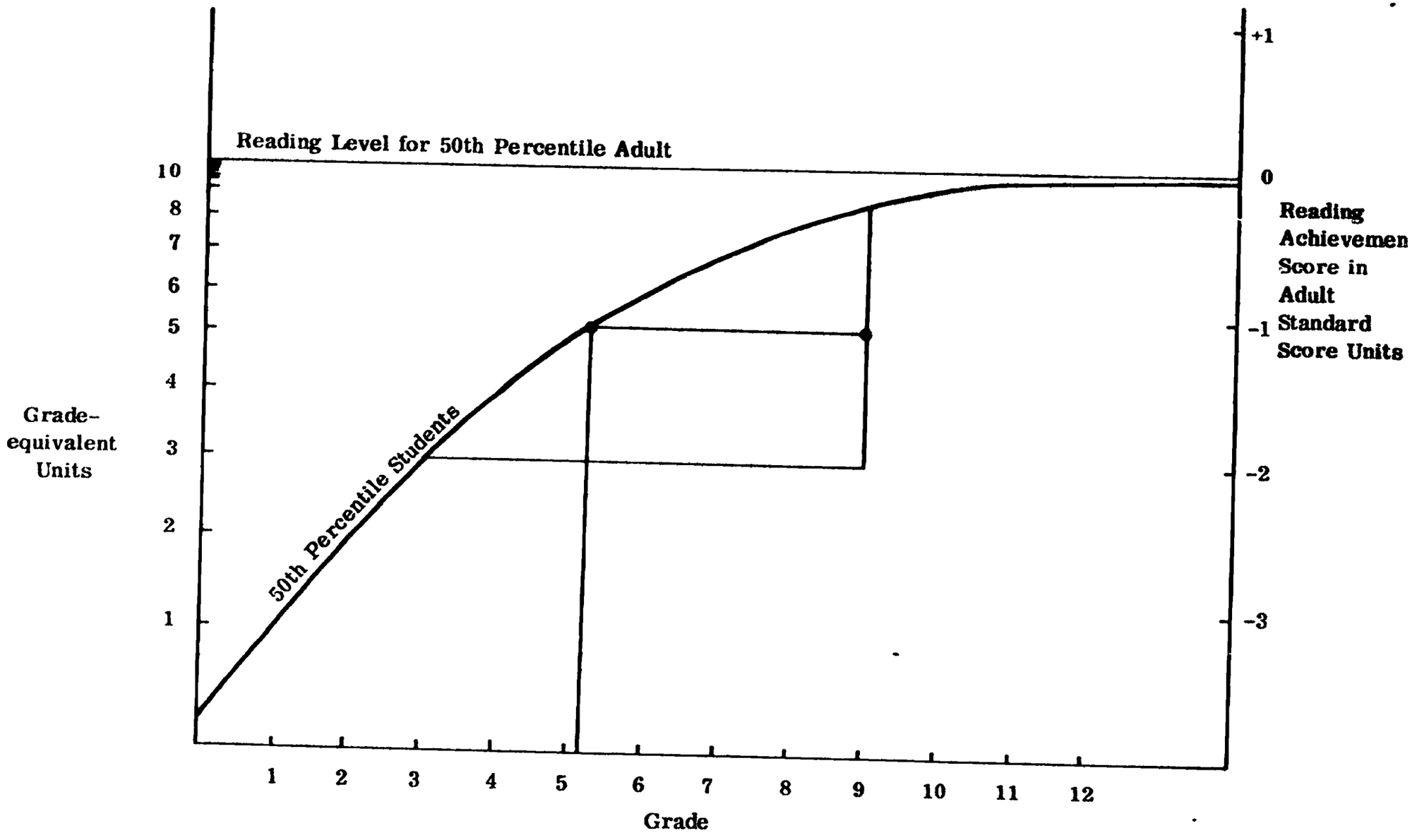


Fig. E2. Hypothetical relationships between grade-equivalent score and reading skill.

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 E1 presents an example of what can happen when scores on a non-equal interval scale are averaged. Two hypothetical students were chosen to represent one standard deviation below the mean and one standard deviation above the mean, respectively, on the Gates-MacGinitie Reading Comprehension Scale. Normative data from grades 6.1 and 6.8 were arbitrarily selected. In this case, using the gain computed from standard scores as the "correct" gain, the mean grade-equivalent score overestimates the true gain by 3.5 months. While the selected example may not be typical with respect to the magnitude of the observed effect, its direction will hold for any negatively accelerated curve, i.e., the shape illustrated in Figure E2.

The second major scaling problem results from the local irregularities in the learning curve which are discussed in detail in Appendix D. The primary cause of these irregularities appears to be the forgetting that occurs over the summer vacation. This phenomenon produces the commonly observed situation in which a class of children achieves lower raw scores on a given test in September than they did the previous June. As illustrated in Figure E3 for example, a single raw score could be the median score for both grades 4.8 and 5.4. While logically, both grade-equivalent scores should be assigned to this raw score, this practice is considered overly confusing, or unesthetic, and is not widely adopted in commercial tests. Instead, some "smoothing" of the data points is done as represented by the solid line in the figure.

The smooth line is used to assign grade-equivalent values to raw scores. This procedure results in a single grade-equivalent value for each raw score but systematically exaggerates the apparent learning gains in experimental situations which use fall and spring testing. For example,

TABLE E1

Mean Scores for Two Hypothetical Students
October and May
Gates-MacGinitie Reading Comprehension Survey D

	Raw Score	Standard Score	Grade-equivalent
Pretest - Grade 6.1			
Student A (84 %ile)	22.50	40.00	3.95
Student B (16 %ile)	46.50	60.00	9.60
Mean	34.50	50.00	6.78
Grade-equivalent	5.40	6.20	6.78
Posttest - Grade 6.8			
Student A (16 %ile)	27.50	40.00	4.55
Student B (84 %ile)	48.00	60.00	10.90
Mean	37.75	50.00	7.73
Grade-equivalent	5.95	6.80	7.73
Grade-equivalent Gain	.55	.60	.95

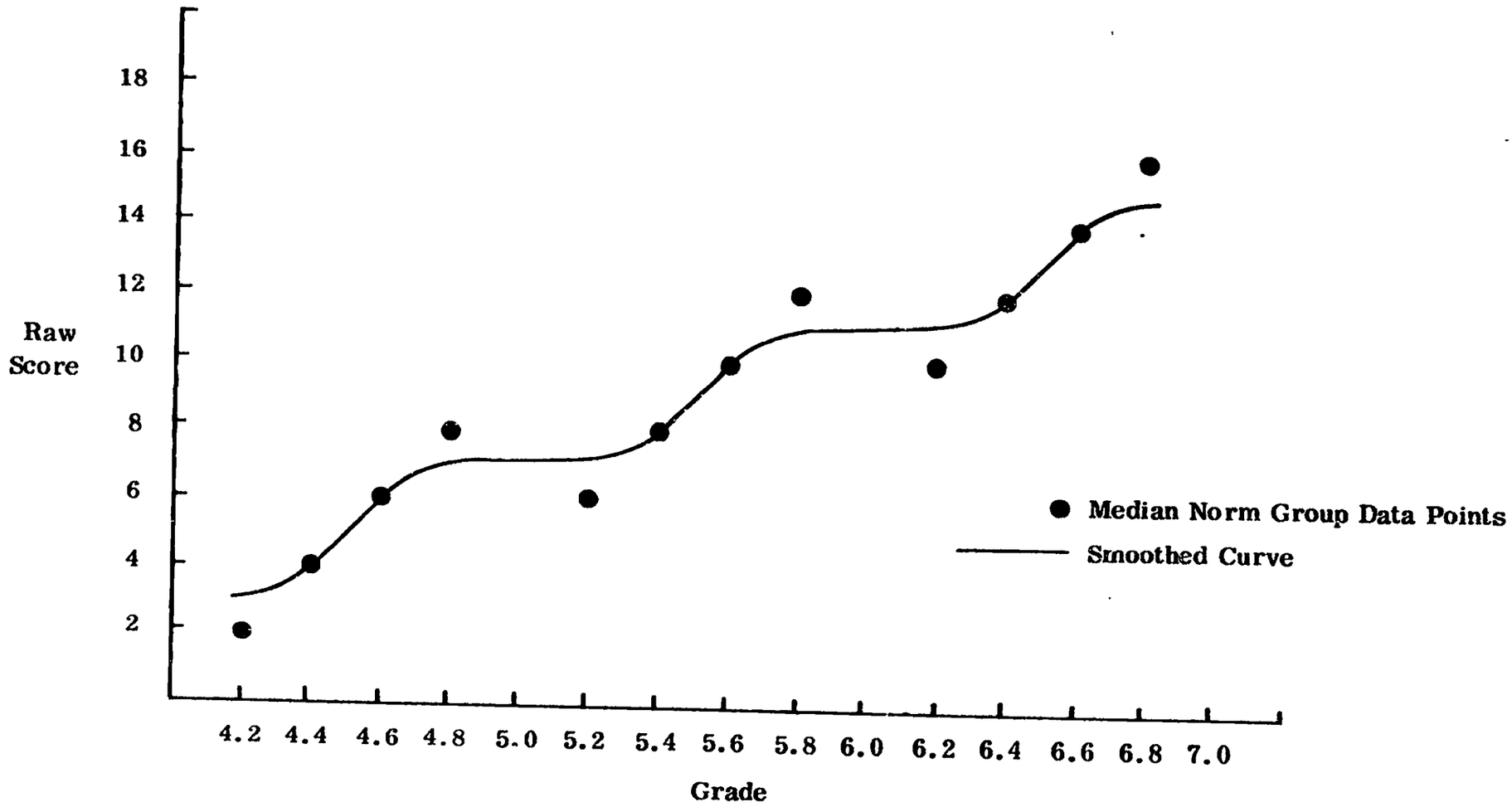


Fig. E3. Hypothetical data points illustrating the process of raw score to grade-equivalent score conversion.

as shown in the figure, a fifth-grade student who scores "6" in November and "12" in May is exactly at the median of his class, but his grade-equivalent scores would indicate that he had progressed from grade 4.6 to grade 6.4--an eighteen-month gain in six months.

The result of using "smoothed" grade-equivalent scores is illustrated graphically in Figure E4. In this figure, the broken line represents the "national norm" in its commonly (mis)conceived linear, month-for-month growth-rate form. The points connected by the solid line are grade-equivalent scores achieved by the median child at each grade level as derived from the smoothed Figure E3 curve. The jagged curve reappears in Figure E4, but in this context it is inherently confusing because, implicit in the concept of grade-equivalent scores, is the notion that the median student's scores "should" fall along the dotted line "national norm." Clearly, they do not, but it is difficult to explain to the uninitiated why the median "grade-equivalent score" for students at grade 4.6 is 5.4, and the grade-equivalent score corresponding to grade 5.2 is 4.6. If a grade-equivalent scores is not, in fact, the score of the median student at that grade level then the interpretation of the score becomes so difficult as to preclude its usefulness. It appears that, in some evaluations, this confusion has led educators to be unduly impressed by very ordinary achievement gains.

It should be noted that this scaling problem is different from the problem of non-comparable test times for norm and experimental groups discussed in Appendix D. Appendix D points out the problems in extrapolating mid-year norm data to fall and spring test dates. The current problem applies to tests which obtain fall and spring norm data but do not accept the data at face value. In both cases the procedure is to artificially smooth an irregular curve, and the effect on project evaluations is to spuriously inflate the apparent amount of learning. It is generally impossible to estimate from information presented in test manuals how much these factors influence test scores or even whether there is any effect at all in specific instances. However, while evidence on the exact magnitude of the effects is sparse, it seems clear that the effects are relatively pervasive.

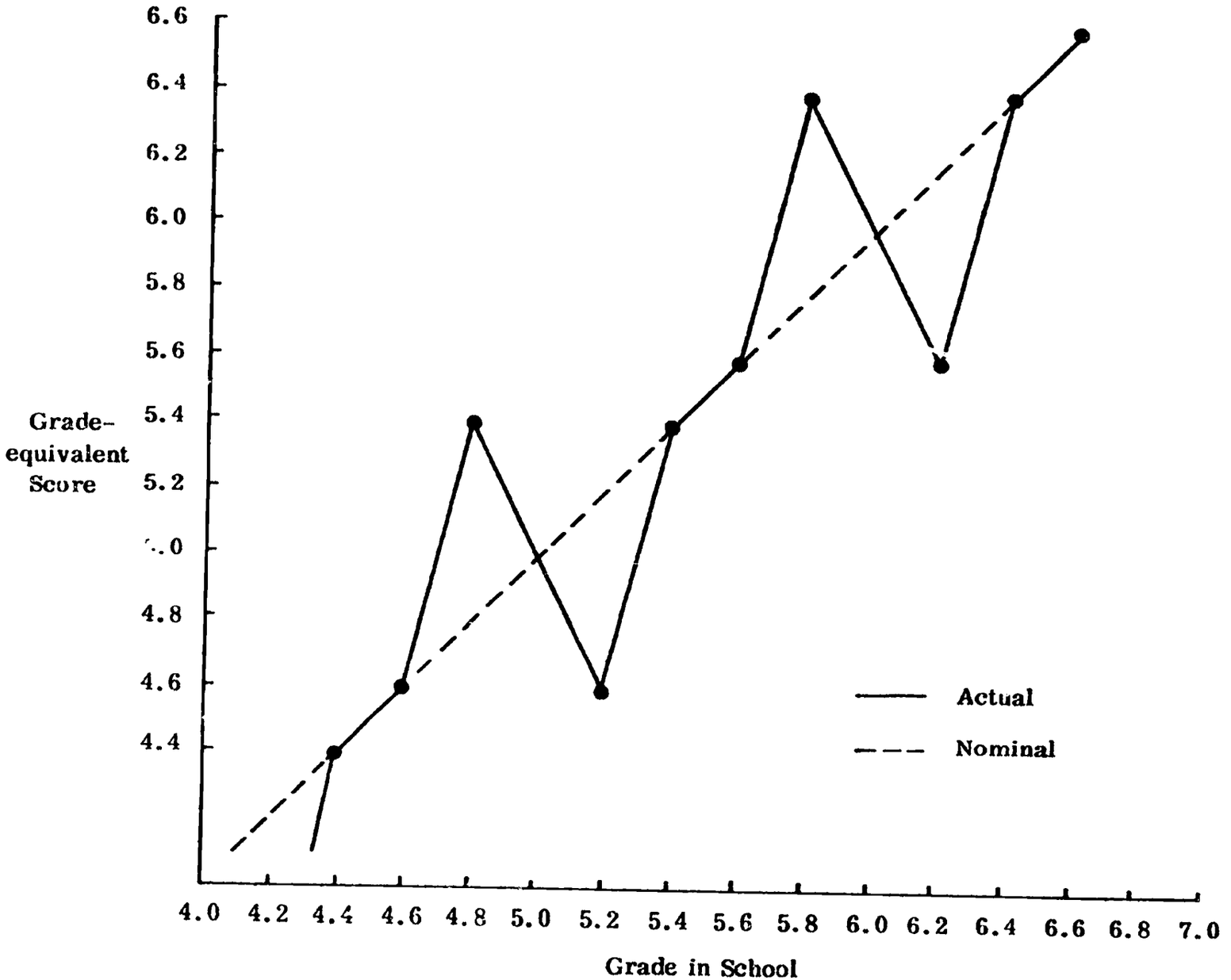


Fig. E4. Hypothetical grade-equivalent scores for 50th percentile students with scores derived from smoothed norm group data.

An additional problem that complicates interpretation of grade-equivalent scores is the restricted range of the typical achievement test. In general, a single test is developed for use in three or fewer grades. Most test companies develop a series of tests of increasing difficulty to cover the entire range of primary (and sometimes secondary) grades. The result is that students scoring more than a year or two "below grade level" may be out of the norm range that was used to develop the test. For example, a test designed for seventh through ninth grade is usually normed on seventh, eighth, and ninth graders. Data may also be collected from sixth and tenth graders. However, the manual may report grade-equivalent scores as low as second or third grade. Obviously, these are simply projected scores since no second or third graders were ever included in the norm group for the test. The error in estimating what median third graders would have scored if they had taken the test is thus added to the problem of interpreting an unequal-interval scale.

Actual data illustrating the above effects are given in Figures E5 and E6. Figure E5 displays grade-equivalent scores for the 16th percentile students (approximately the mean of the bottom quartile). The scores were taken from the manual of a widely used reading test. They were derived from normative data collected by the test developers and reflect the same type of data as the hypothetical smoothed curve in Figure E3 except that the vertical axis is scaled in grade equivalents rather than raw scores. The data have been smoothed, according to the accompanying technical manual, but the extent of the smoothing is not reported.

It will be noted that within-year gains are, in general, closer to month-for-month than are between-year gains. We cannot tell from the reported information to what extent (if any) the smoothing has reduced this effect. It is important to keep in mind, however, that the only reason the effect is observable at all is that the test in question includes normative data from two points in the school year: October and April. The reported norms which are plotted in Figure E5 (October,

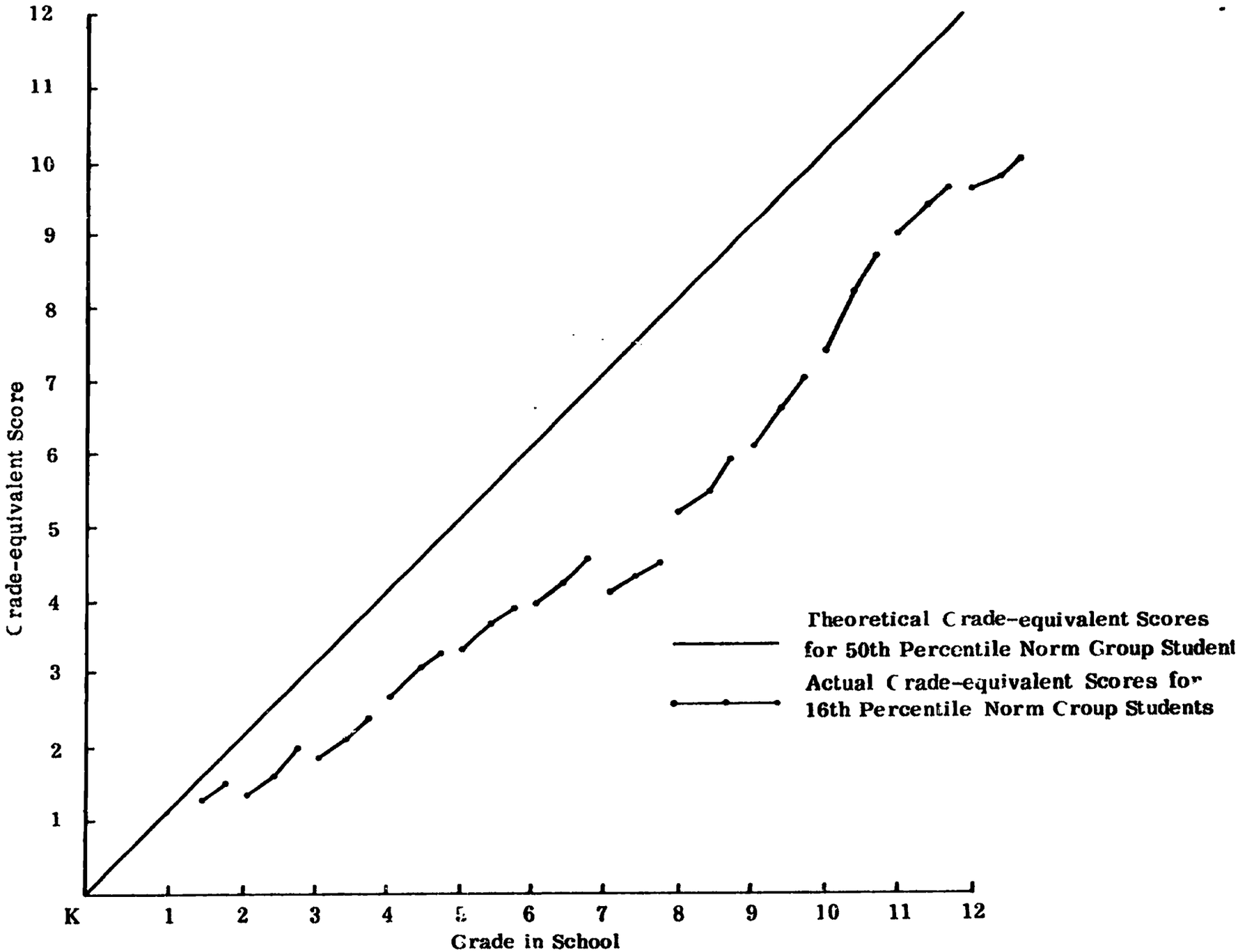


Fig. E5. Assignment of grade-equivalent scores for 16th percentile norm group students, Gates-MacGinitie Reading Comprehension Tests.

February, and May) are extrapolated from the two actual test dates. Many widely used tests preclude the detection of any discrepancy between within-year and between-year learning rates by collecting normative data at only one time during the year and extrapolating to provide intermediate "norm data" (see Appendix D).

It will also be noted that the appearance of the curve changes after grade six. It is not clear what produces the change but it seems likely that one factor is the relatively high drop-out rate of low scoring students in junior and senior high school. The sixteenth-percentile student in the high-school norm group probably stood relatively much higher in his first-grade peer group distribution simply because first-grade distributions include a large number of slow students who drop out before reaching high school.

Figure E6 presents data from a study by Tallmadge (1973) of all California Title I students. This curve is analogous to the schematic curve illustrated in Figure E4. It is based on a variety of tests and includes the effects of both smoothing and non-comparable norm times. These effects are undoubtedly confounded with those of other extraneous variables, as Tallmadge points out:

> There is some danger in interpreting Figures 1 and 2 as if they represented longitudinal data. They do not--the data are cross sectional and each year's growth is represented by a different sample of pupils. For this reason it is not strictly legitimate to talk about losses over the summer. We do not know how those children represented by each pretest point on the figures scored at posttest time the year before. Still, it seems reasonable to assume that many, and perhaps most, of the children served by Title I in the sixth grade this year were also served last year in the fifth grade and in earlier grades and years as well. Until data are acquired over at least a 12-month interval (ideally from posttest one year to posttest the following year), questions of this sort must remain unanswered.

Hopefully, it is clear from the above discussion that the apparent simplicity of grade-equivalent scores obscures their basically complex nature. While they may serve some purpose in individual counseling and

ERIC

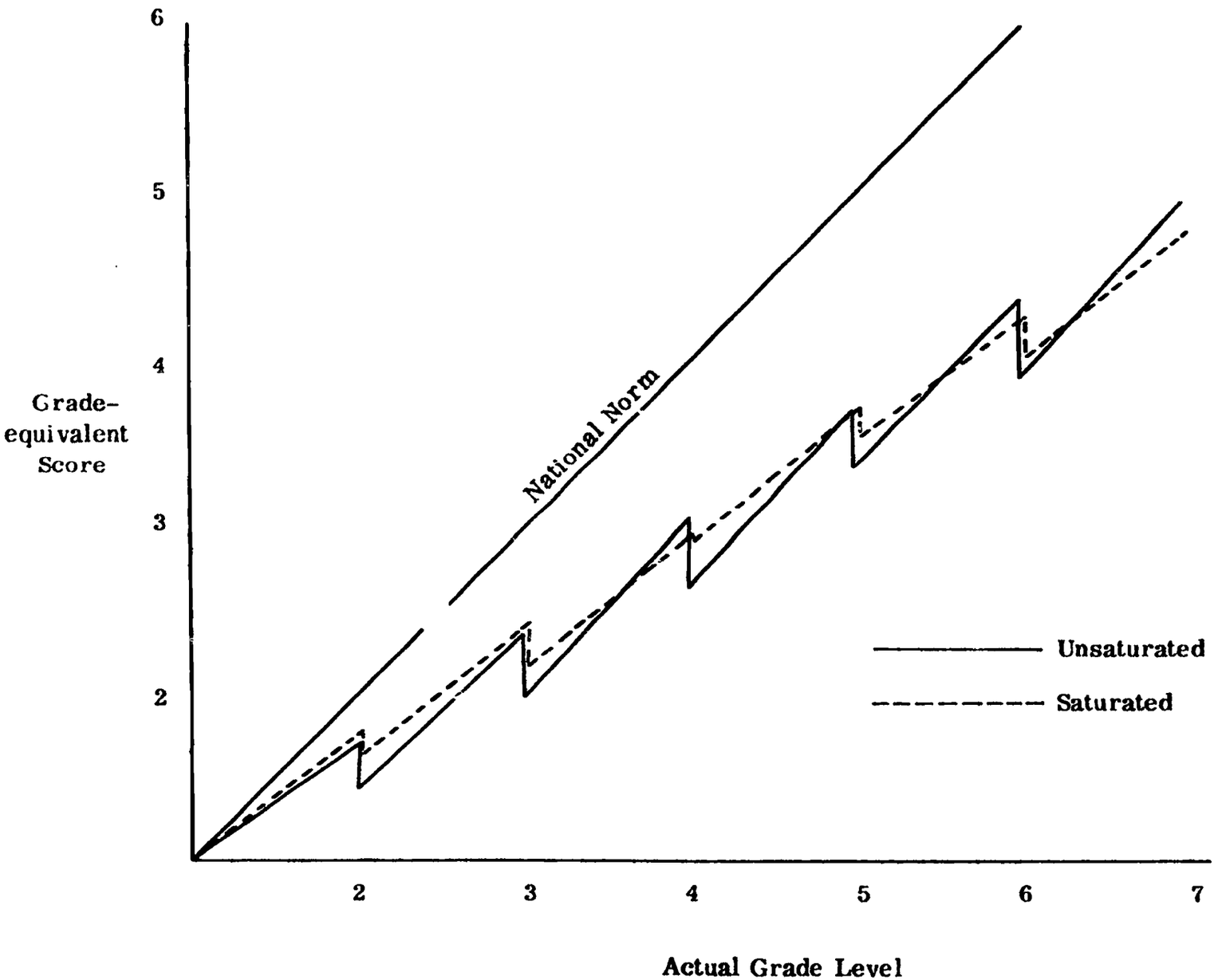


Fig. E6. Reading growth rates in saturated and unsaturated schools.

guidance, the purpose for which the achievement tests were designed, the current widespread use of grade-equivalent scores in evaluating educational programs can only be considered extremely unfortunate.

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.

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., & 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.

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

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

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), Psychological principals 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. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

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

Horst, P. Effect of treatment as a special case of generalized multiple regression. Research Bulletin. Eugene, Oregon: Oregon Research Institute, 1974.

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, 1963.

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

Parsons, H. M. What happened at Hawthorne? Science, 1974, 183, 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.

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.

Tallmadge, G. K. An analysis of the relationship between reading and mathematics achievement gains and per-pupil expenditures in California Title I projects, fiscal year 1972. Palo Alto, Calif.: American Institutes for Research, March 1973. (AIR-35100-3/73-FR)

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

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