DOCUMENT RESUME

ED 259 009 TM 850 378

AUTHOR Madhere, Serge
TITLE Efficiency Analysis: Enhancing the Statistical and Evaluative Power of the Regression-Discontinuity Design.
PUB DATE Apr 85
NOTE 24p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Decision Making; Efficiency; Elementary Education; Elementary School Mathematics; *Evaluation Methods; *Evaluation Utilization; Hypothesis Testing; *Mathematical Models; Pretests Posttests; *Program Evaluation; Regression (Statistics); Remedial Mathematics; Research Design
IDENTIFIERS *Efficiency Index; Evaluation Problems; Regression Discontinuity Model

ABSTRACT

An analytic procedure, efficiency analysis, is proposed for improving the utility of quantitative program evaluation for decision making. The three features of the procedure are explained: (1) for statistical control, it adopts and extends the regression-discontinuity design; (2) for statistical inferences, it de-emphasizes hypothesis testing in favor of interval estimation; and (3) it uses the limits of the confidence interval to qualify the level at which a program operates, rather than making a simple statement about goal attainment. Application of this procedure is illustrated with data obtained at grades 2, 3, 7, and 8 for a remedial mathematics program. Methods are suggested for making these evaluation results accessible and practical for educational decisionmakers through graphic presentations and by linking evaluation outcomes to particular administrative objectives, specific planning procedures, and a set of corrective and/or supportive program activities. (Author/GDC)

Efficiency Analysis:

Enhancing the Statistical and Evaluative Power of the Regression - Discontinuity Design

Serge Madhere, Ph. D.

Newark Board of Education
Division of Research and Evaluation
2 Cedar Street
Newark, New Jersey 07102

April 1985

Efficiency Analysis:
Enhancing the Statistical and Evaluative Power of the Regression-Discontinuity Design

Abstract

This study describes an analytic procedure that aims at improving the utility of quantitative program evaluation for decision-makers. The procedure has three main features: a) For statistical control, it adopts and extends the regression-discontinuity design. b) For statistical inferences, it de-emphasizes hypothesis testing in favor of interval estimation. c) It uses the limits of the confidence interval to qualify the level at which a program operates, rather than making a simple statement about goal attainment. Following a step-by-step illustration of the quantitative procedure, we show how each type of evaluation outcome thus obtained can be linked to a particular administrative objective and/or orientation, some specific planning procedures, and a set of corrective/supportive program activities.

Efficiency Analysis:

Enhancing the Statistical and Evaluative Power of the Regression - Discontinuity Design

Introduction

In their review of the evaluation process, Stufflebeam, Foley, et al. (1971) discussed five shortcomings that greatly limit the value of evaluation to decision-makers in their effort to improve an educational program. These five shortcomings are: a) the poor linkage between educational theory and evaluation practices; b) a lack of appropriate designs or even of instruments for the evaluative tasks; c) the shortage of personnel with a working knowledge of both evaluation techniques and the decision-making process; d) the narrowness of quantitative criteria which, too often, lead to the improper conclusion of no significant difference; e) the esoteric nature or poor quality of the information generated through the evaluation.

Some of the shortcomings have, since then, been addressed. For instance, Tallmadge, Horst, and Wood (1975) have adapted and publicized three quasi-experimental models to guide the assessment of project impact on student achievement. Strenio, Weisberg, and Bryk (1979) have offered a model of cognitive growth that can be applied in different evaluation contexts. (See also Keats, 1983.) Stufflebeam et al. (1971) have shown the relevance of the work of Braybrooke and Lindblom (1963) for making evaluation results congruent with the administrative decision-making process. But evaluators are still grappling with some of these issues: Should the quantitative approach to evaluation, with its limited focus on program outcomes, be replaced by an observational, ethnographic approach? If not, is the hypothesis testing paradigm, so valued in experimental research, appropriate for program evaluation? How can quantitative information be accurately translated into terms that are understandable by educational managers?

This paper represents a modest attempt to deal with the last two questions. It described a strategy developed over the past three and a half years for the evaluation of a compensatory education program in an inner-city school district. The strategy, termed efficiency analysis, takes on the following features: a) For statistical control, it makes use of the regression-discontinuity design. b) For statistical inferences, it de-emphasizes hypothesis testing in favor of interval estimation. It uses the boundaries of the confidence interval to describe the level at which the program operates, rather than making a simple statement about goal attainment. c) It translates the quantitative description into an unequivocal decision alternative for the program administrators.

Evaluation Design

The regression-discontinuity (Campbell and Stanley, 1963) is a quasi-experimental design appropriate for situations where there is a known interaction between treatment assignment and ability (achievement, aptitude, etc.). It has emerged in recent years as one of the most promising quantitative models for the evaluation of compensatory education. Based on the criterion of internal validity, the regression-discontinuity design has been shown to be superior to the norm-referenced model (Linn, 1981), since there often are multiple academic and contextual differences between the remedial group under study and the national sample from which test norms are developed. Based on the criterion of feasibility, the regression-discontinuity design has been found preferable to the classical experimental/control group approach, since it is impractical or unethical, in many instances, to withhold needed services fr. students in order to set up a comparison group (Wolf, 1981). Beyond the issue of applicability, the design may be most desirable, 1) when assignment to the 'treatment' group is based on a definite

cutoff score, i. e., all students with a pretest score below a certain mark participate in the remedial program, while those above are dispensed of it; 2) when the educational environment includes multiple 'treatments,' and there is a need to separate the impact of the remedial, supplementary intervention from that of the general program of instruction. To determine the treatment's effectiveness, the task of the evaluator is to estimate what the performance level of the low achieving group would be without the remedial support, then, one test to see whether the actual score for that group is significantly different from the expected value.

Two variants of this design exist. In the strict regression-discontinuity approach, separate pretest-posttest regression lines are obtained for the group above and the group below the cutoff point. The, two predicted values for that pretest cutoff score are calculated, by fitting it into each regression equation. A discontinuity in the regression lines, i.e., a difference between the predicted cutoff values, if significant, is taken as a measure of program impact. Tallmadge, Horst, and Wood (1975) propose a modification of the original technique that may be more sensitive to a possible pretest/program interaction among the low achieving students. In this version, known as regression-projection, the relationship between the pretest and the posttest is calculated only for the group of students above the cutoff score. Then, assuming linearity over the entire range of pretest scores, a single regression coefficient is used to estimate what the remedial group's posttest mean would have been under a 'no-treatment" condition. The formula for making such an estimate reads as:

$$E(\bar{Y}_t) = \bar{Y}_c + b_c(\bar{X}_t - \bar{X}_c)$$

[Insert Figure 1 here]

It simply means that the difference between the high achieving and the low achieving group on the posttest is expected to be the same as it was on the pretest, except for the imperfect correlation between the two measures. Any discrepancy between the projected and the observed posttest mean is attributed to the remedial treatment. The two versions of the regression design are illustrated in Figure 1. The details of the statistical test to establish significance of the differences can be found in Sween (1971) for the regression-discontinuity, and in Tallmadge and Horst (1974) for the regression-projection.

Statistical Analysis

The statistical tests offered to accompany the regression designs aim at 'proving' a single point: that the program has or has not met its objective. As such, they follow the hypothesis testing paradigm, which is the one most commonly used in psychological and educational research. But, hypothesis testing is only one means of deriving statistical inference. As stated by Hays (1963), "in many circumstances," (and evaluation seems to be exactly one of these circumstances) "the primary purpose of data collection is not to test a hypothesis, but rather to obtain an estimate of some parameter" (p. 375). A range of values may be more useful or more stable than a single, unqualified estimate, given the presence of sampling error affecting most research data. Rather than just ignoring the sampling error, an evaluator can place him/her self on safer ground by dealing straight forwardly with it, when drawing a conclusion about program effectiveness. To do that, one can turn to another form of statistical inference, the calculation of a confidence interval.

Ordinarily, in regression analysis, it is possible to establish confidence intervals for three different parameters: the regression coefficient itself, the actual score of an individual on the criterion measure, or the predicted value of a particular pretest score. Given the critical role accorded to the predicted mean value or to the cutoff in the regression design, the calculation of the confidence interval is most necessary for each of these parameters. To obtain the boundaries of the confidence interval, one can use the following formula adapted from Hays (1963):

$$Y'_t \pm (t_{\alpha/2})\,(\text{est}\,\sigma_{yx})\sqrt{\frac{1}{N} + \frac{(\bar{X}t - \bar{X}c)^2}{NS_x^2}}$$

where: Y'_t = Predicted posttest mean for the treatment group, or the predicted posttest value for the cutoff score

X_t = Mean of the treatment group on the pretest, or the cutoff score on the pretest

$\bar{X}_c$ = Mean of the control group on the pretest

$\text{est}\,\sigma_{yx}$ = The standard error of estimate adjusted by the sample size

$$\text{est}\,\sigma_{yx} = \sqrt{\frac{NS^2\,(1 - r^2)}{N - 2}}$$

For the t-value, any probability may be retained by the evaluator, depending on the desired level of confidence interval. For a 95% confidence interval, t is set at 1.96.

Two kinds of information can be derived from such an analysis. One kind pertains to the total change in students' classification or the proficiency rate of the program; the other concerns the relative amount of gain achieved by students in the program.

A - Success or Proficiency Rate

When the confidence interval is calculated for the predicted cutoff value on the posttest, its upper limit indicates the highest possible score that one

would a priori expect for a participant in the remedial program. By inspecting the score distribution one can, then, determine the percentage of participants scoring above that mark. Those are students who have made so much progress that they are no longer in need of remediation. Their percentage is likely to be small; but, it is a clear indication of a program's impact, and one that is readily understood by administrators. We call this percentage the proficiency rate yielded by the program.

B - Efficiency Level

When one turns the focus on the predicted mean value, one can obtain some additional and finer reference points to describe the program. If the actual posttest mean for the treatment group does not fall within the calculated interval, one can be 95 percent confident that 'something extraordinary' is happening with the program. If the observed mean is above the upper limit of the confidence interval, the impact of the program is definitely positive. On the other hand, if the observed mean is below the lower limit of the confidence interval, the return on the program is clearly not what one would expect. As one can see, the procedure is quite unequivocal about the extreme cases. One may say that it also increases the likelihood of arriving at a nonsignificant difference. But even within the region of nonsignificance, it is possible to set up a gradient of performance, which allows the evaluator to draw inferences not just about goal attainment, but also the level at which a program operates. Indeed, all the bits of information obtained from the standard statistical analysis can be condensed into one measure that we call the efficiency index. The term efficiency speaks of the average amount of progress made by the treatment group participants, relative to their own entry level and that of students in the control group. Mathematically, it is calculated according to the following formula:

$$F = \left[1 - \frac{\bar{Y}_c + \bar{Y}_t + b\,(\bar{X}_t - \bar{X}_c)}{t_\alpha\,(\text{est}\sigma_{yx})\sqrt{\frac{1}{N} + \frac{(X_t - X_c)^2}{NS_X^2}}}\right] * .5 = \left[1 - \frac{2\,(\bar{Y}'_t - \bar{Y}_t)}{R}\right] * .5$$

where: R = the range of points over the confidence interval

$\bar{Y}_t$ = mean on the posttest for the treatment group

$\bar{Y}'_t$ predicted posttest mean for the treatment group

If the observed and the predicted posttest means coincide, the efficiency index will take the value of .5. If the observed posttest mean corresponds exactly to the upper limit of the confidence interval, the efficiency index will take the value of +1. If the observed posttest mean falls precisely at the lower boundary of the confidence interval, the efficiency index will take the value of 0.

Although the derivation of such an index may seem complex, its merit is that it tremendously simplifies the reporting of evaluation results to program administrators. That advantage can be appreciated when one has to deal with a program implemented at several grade levels. Whenever the efficiency index is greater than 1, the program is probably exemplary; whenever the efficiency index is negative, the program is probably in trouble. Even when the index falls between 0 and 1, (in other words, no statistical significance is obtained), it is still possible to call attention to different degrees of efficiency; in that sense, the procedure gets around the no-significant difference symptom that Stufflebeam et al. complained about.

The whole procedure is illustrated below with actual data obtained at four grade levels (2, 3, 7, and 8) for a remedial math program.

In grade 7, for example, students with a pretest score lower than 38 NCEs (29th percentile rank) were assigned to the remedial program. The average pretest score for this low achieving group was 30.64 NCE, compared to a mean

of 57.49 for students not participating in the program. Based on the regression analysis, it was projected that the posttest performance for students in the first group would be around 25.8 NCE, in the absence of the remedial program.

$$\bar{Y}'_t = 55.03 + .77 \frac{17.00}{12.01} (30.64 - 57.49) = 25.78$$

A 95 percent confidence interval was calculated, that extends ± 6.90 NCE points around that central value.

$$25.78 \pm (1.96)(11.03)\sqrt{\frac{1}{59} + \frac{(30.64 - 57.49^2}{59 \times (12.01)^2}} = 25.78 \pm 6.90$$

The observed posttest mean for the treatment group was 34.02, and fell outside of the confidence interval. It actually exceeded its upper limit by 1.34 NCE. That difference can be translated into an efficiency index equal:

$$E = 1 - \frac{2\ (25.78 - 34.02)}{13.8} * .5 = 1.10$$

Clearly, the impact of the program is strongly positive at that grade level, for the average participating students. It is desirable to determine how many of them will no longer need remedial support. The regression analysis led to a projected score of 33.79, corresponding to the pretest cutoff of 38 NCE.

$$Y'_{c.o} = 55.03 + .77 \frac{17.00}{12.01} (38 - 57.49) = 33.79$$

A 95 percent confidence interval was also estimated for that value, and its upper limit turned out to be 39.17 NCE, (33.79 + 5.38). A study of the posttest score distribution revealed that 33 percent of the participants achieved above that level. Similar calculation can be carried out for each grade.

[Insert Table 1 here]

Management Information

Two questions need to be addressed now: 1) How does one convey that kind of complex information to administrators in a handy way? 2) How does one advance the probability that the reported information indeed be included in the decision-making process?

^ - Making it Accessible

The time-honored way of conveying a great deal of quantitative information in a handy and attractive way is through graphics. It is at this point that evaluation is no longer a science, but becomes an art. The evaluator must be resourceful, and the graphic capabilities of microcomputers are now available to enhance that resourcefulness. One can use three types of graphs to summarize the information obtained through the regression-discontinuity design: a) Information on the success or proficiency rate of a program may be reported on a bar graph, as illustrated in Figure 2. b) Information on a

[Insert Figure 2 here]

program's efficiency may be reported in a modified scattergram as follows. The horizontal axis shows the pretest scores (say in NCE's) with a clear mark for the cutoff point; the vertical axis shows different values of the efficiency index. One can divide the area delineated by these axes into three subfields, by drawing two lines at point 1 and 0, perpendicular to the efficiency axis. The top line, at point 1, corresponds of course to the upper limit of the confidence intervals calculated; it can be referred to as the optimal efficiency line. The bottom line, at point 0, corresponds to the lower limit of the confidence intervals calculated; it may be referred to as the minimal efficiency line. The subfield above the optimal efficiency line is designated as a net growth area; the subfield between the optimal and the minimal efficiency lines is designated as a maintenance area; the subfield below the minimal efficiency line is designated as a breakdown area. The

points in the scatterplot represent the various sites or grade levels at which the program was implemented. If at a particular grade level the actual posttest mean falls within the confidence interval, for the predicted mean, that observation will appear between the two efficiency lines; this will suggest that the remedial program is operating as a maintenance unit, whose utility is to prevent the deteri ration of sk ls, and thus sustain the operation of the regular instructional program; in other words, without it, the regular program of instruction may not be able to function with any kind of efficacy. If at another grade level the posttest mean exceeds the upper limit of the confidence interval, that observation will appear above the optimal efficiency line; this will suggest that the remedial program is operating as a producti unit, capable of creating a net growth in students' competence. If at still another grade level the posttest mean fails to reach the lower limit of the confidence interval, that observation will appear below the minimal efficiency line; this will suggest that the remedial program is in disrepair. The whole procedure for reporting information on program efficiency is depicted in Figure 3. c) The two kinds of information on

[Insert Figure 3 here]

efficiency and success/proficiency rate can be integrated in one diagram, called a performance record. As shown in Figure 4, each grade level is

[Insert Figure 4 here]

represented at the center of the diagram. The measures of program performance are indicated numerically at the periphery, and graphically as grooves on the record. The inner marks stand for the degree of program efficiency, while the outer marks stand for proficiency. These three types of graphs can be attached to the Executive Summary for the evaluation report.

B - Making it Practical

In order to make the information he/she generates relevant to the decision-making process, the evaluator must have a good understanding of that

process. The understanding should be based on empirical evidence about the overall program environment, and should also be be guided by a theoretical framework. Previous research suggests that the process of rational decision-making follows four principles. What are those principles and what do they entail?

1. A decision requires a clear information base.

The information base, which is of course nothing other than previous evaluation results, may indicate one of three things: a) a given program is capable of producing net academic growth, i. e., its efficiency index is greater than 1; b) a given program operates as a maintenance unit, i. e., its efficiency index is between 0 and 1; c) a given program is experiencing a breakdown, i. e., its efficiency index is lower than 0.

2. A decision is always inscribed within a general approach to management.

Following Stufflebeam et al. (1971), we distinguish three possible approaches in an educational setting: a) a homeostatic approach, intended to sustain the achieved balance in a program; b) an incremental approach, aimed at "shifting the program to a new balance based upon small serial improvements" (p. 69); c) a neomobilistic approach geared for a large and significant change necessitated by critical program conditions.

3. A decision calls for selection or design of specific procedures to be followed.

This principle really speaks of the planning stage in the process. a) Planning may consist in simply standardizing or operationalizing the procedures presently in use. b) Another possibility is to target particular areas where the need is the greatest, or where resource allocation will be most efficient. c) Still another alternative is to reorganize a program in all its aspects, adjusting the objectives, providing new means, redefining personnel

roles, setting check points for accountability.

4. A decision involves translating a set of selected procedures into activities in order to meet an objective.

Three courses of action may be followed: a) one can continue or recycle a set of practices proven to be successful; b) one can offer training and other activities in staff development; c) one can move to enforce or implement available guidelines/procedures where numerous discrepancies have been found between a program's objectives and modus operandi.

Stufflebeam et al. insist that the ultimate objective of a rational decision-making process, similar to the one outlined above, is educational improvement. While no educator would contest that view, it has been our experience that a number of immediate goals often supersede the ultimate objective. These immediate administrative goals fall into three categories: those aimed at producing change, those aimed at achieving control, those aimed at promoting or marketing a particular program or position for public relations purposes. These immediate goals, because of the rather quick payoffs associated with them, are the guiding lights of management. So, the evaluation results must be articulated to them in order to sensitize the decision-makers. We propose a restructuring of the decision-making model to reflect that situation. Figure 5 depicts this new structure.

The model establishes a correspondence between each immediate goal and the type of elements in the decision-making process which it seems most congruent with. It can be of great utility to the evaluator in formulating his/her recommendations for program development. Depending on the kind of evaluation results obtained (i. e., the value of the efficiency index), a particular administrative approach, some specific planning procedures, and a set of corrective/supportive activities may be suggested. That kind of

detailed, facilitative work has a good probability of catching the attention of the decision-makers.

Table 1

Statistical Data for Chapter I and Nonchapter I Students in Mathematics

Parameters \ Grade	2		3		7		8	
	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.
1. Pretest Mean	32.04	64.80	23.27	60.00	30.64	57.49	30.09	56.83
2. SD of Pretest	11.26	14.89	9.91	16.73	8.31	12.01	9.36	14.46
3. Posttest Mean	37.70	58.94	32.98	59.13	34.02	55.03	37.40	56.02
4. SD for Posttest	17.27	19.39	10.95	16.31	11.88	17.00	8.15	14.58
5. Cutoff Score	41.90	-	28.20	-	38.00	-	38.00	-
6. Pre-Post Correlation	-	.57	-	.39	-	.77	-	.59
7. Sample Size (N)	70	65	64	61	58	59	66	60
8. Expected Post Mean	34.75		44.70		25.78		40.12	
9. Confidence Interval for (8)	±9.48		±10.06		±6.90		±6.53	
10. Expected Value for Cutoff	44.29		47.04		33.79		44.82	
11. Efficiency Index	+.65		-.08		+1.10		+.29	
12. Proficiency index	17%		2%		33%		5%	

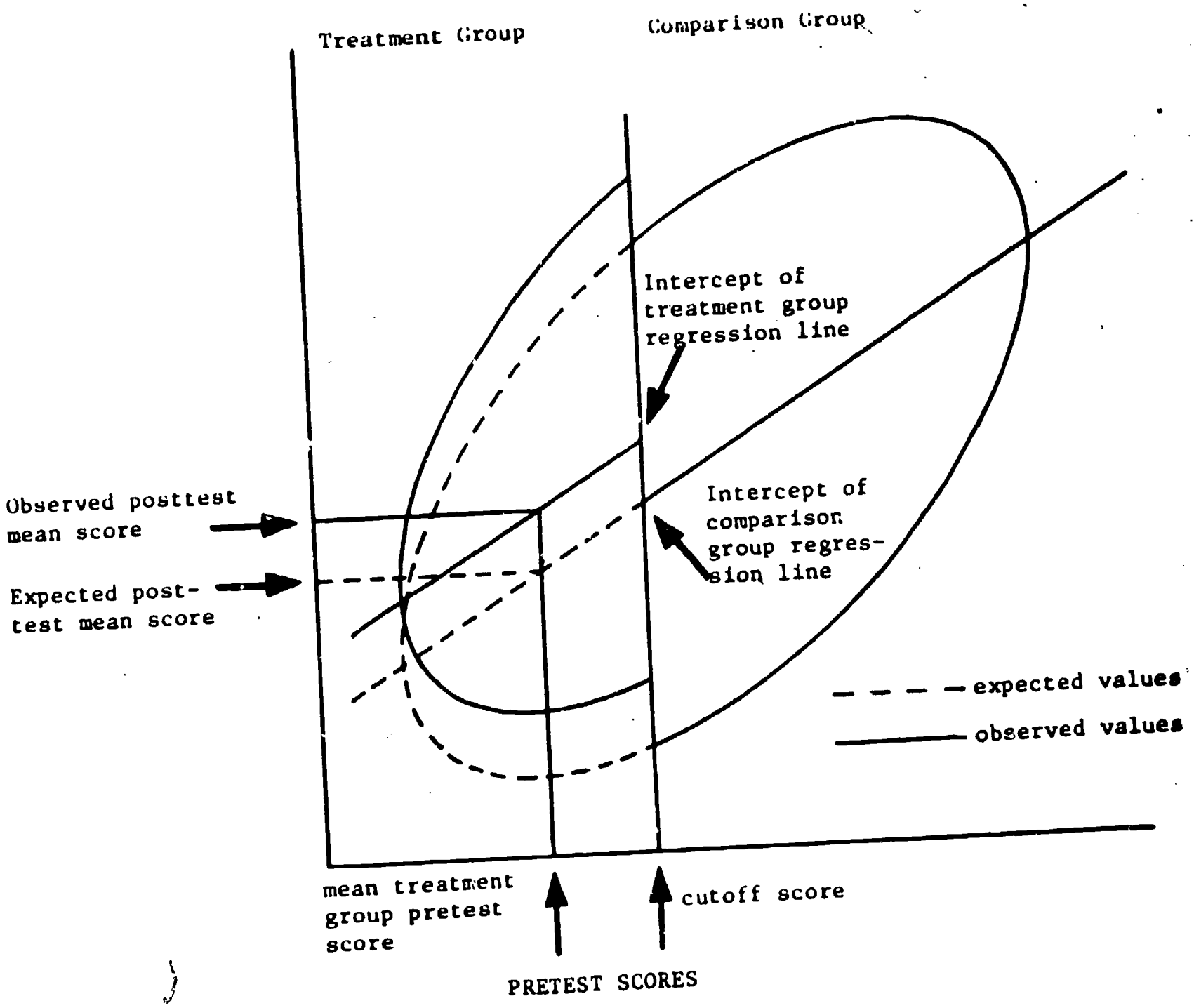

Fig. 1. Score distributions with treatment effect independent of pretest status. (reprinted from Tallmadge and Horst, 1976)

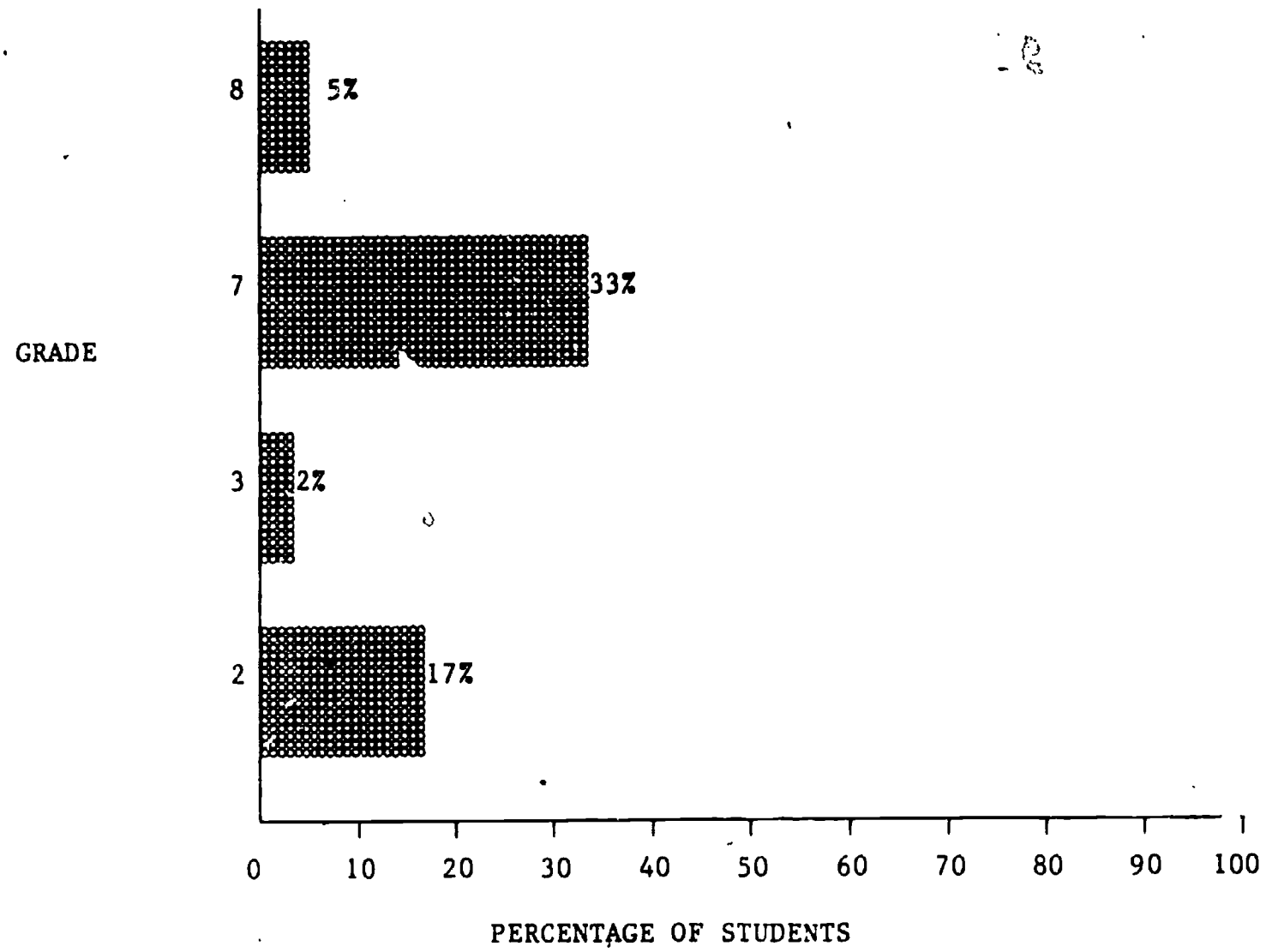

FIGURE 2 - PROFICIENCY DATA: PERCENTAGE OF STUDENTS 'GRADUATING OUT' OF THE REMEDIAL PROGRAM AT EACH GRADE LEVEL SERVED

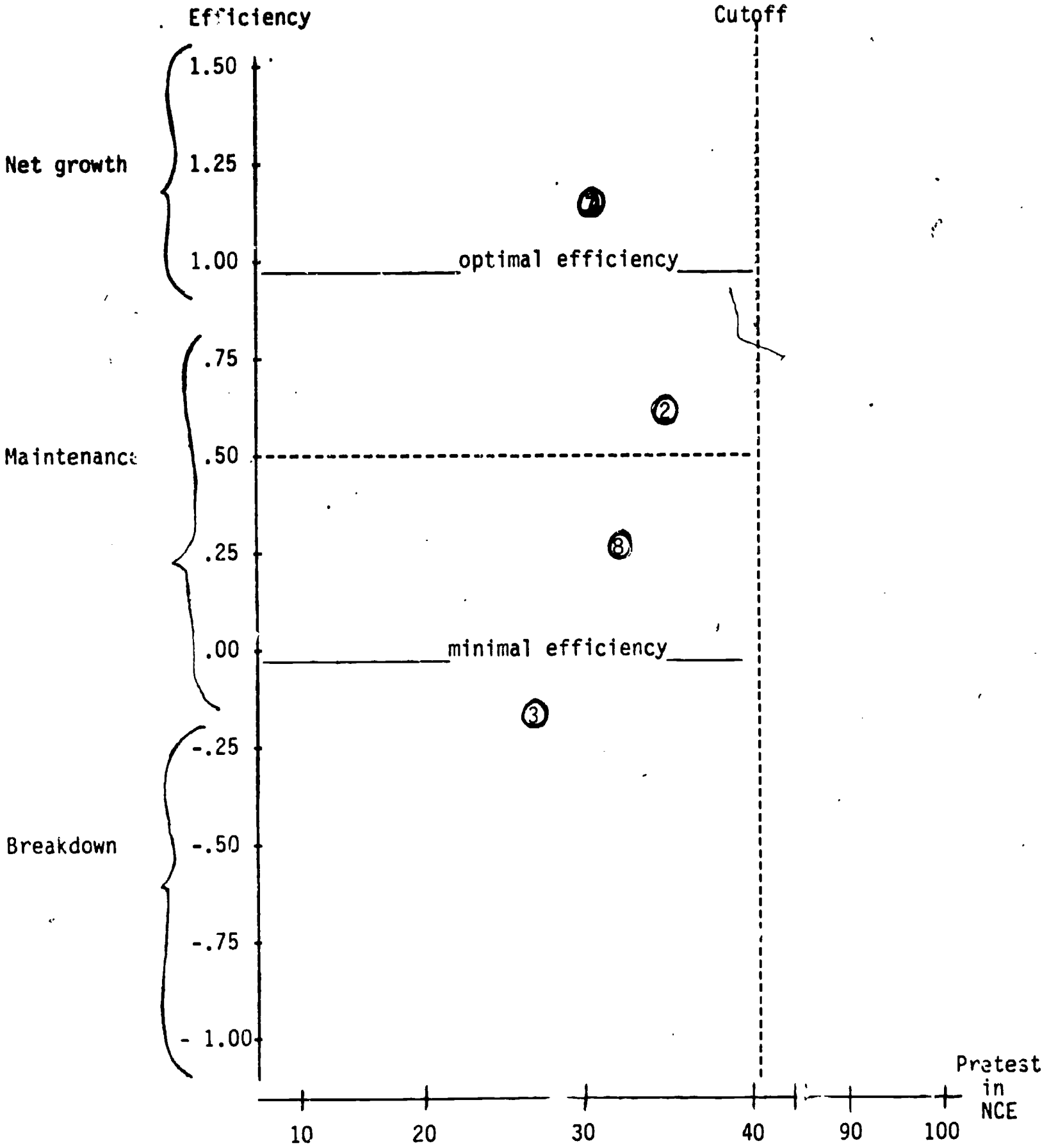

FIGURE 3 - PROGRAM EFFICIENCY LEVEL AT FOUR GRADES

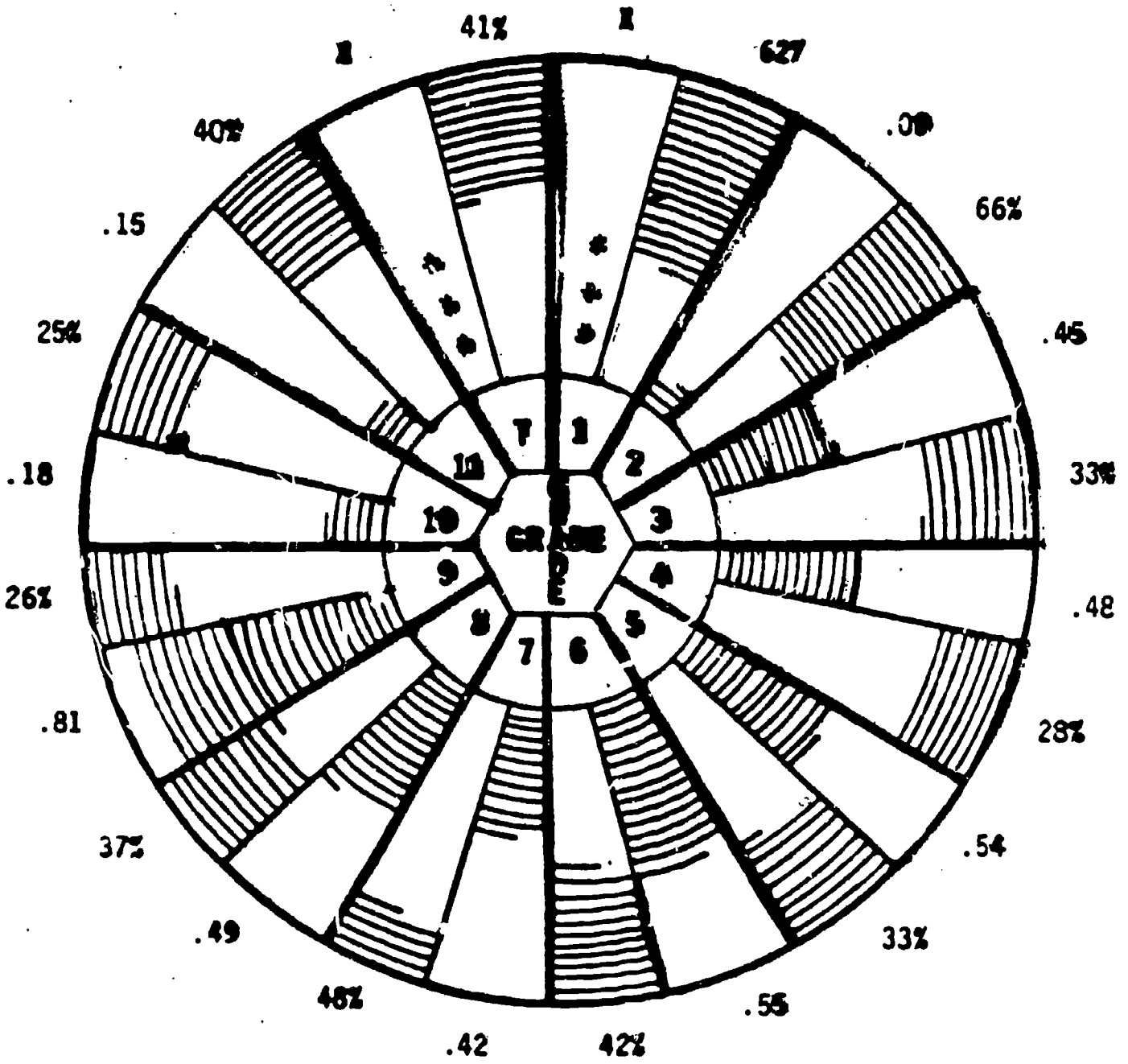

Figure 4 - Math Record

(Each line is worth 4 percentage points. Inner marks represent the degree of program efficiency. Outer marks represent the degree of program effectiveness).

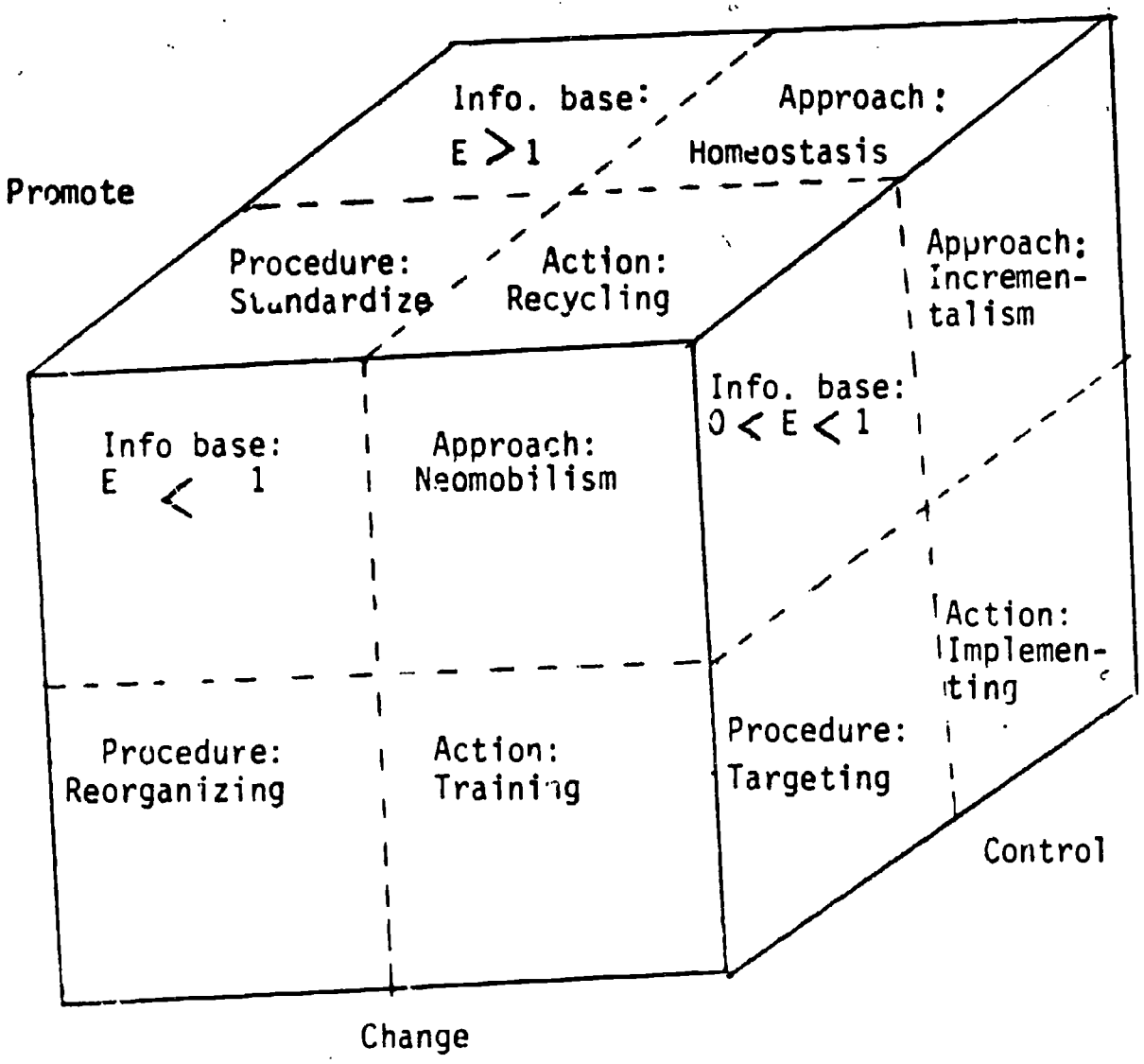

FIGURE 5 - STRUCTURE OF RATIONAL DECISION-MAKING

References

Braybrooke, D. and Lindbloom, C. E. A Strategy of Decision. New York: Free Press, 1963.

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1963.

Hays, W. L. Statistics for Psychologists (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Keats, J. A. Ability Measures and Theories of Cognitive Development. In. H. Wainer and S. Messick (eds.) Principles of Modern Psychological Measurement. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1983.

Linn, R. L. Measuring Pretest-Posttest Performance Changes. In R. A. Berk (ed.): Educational Evaluation Methodology. Baltimore: John Hopkins University Press, 1981.

Strenio, J. F. Weisberg, H. I., and Bryk, A. S. Empirical Bayes Estimations of Individual Growth Curve Parameters. Cambridge, Massachusetts: Huron Institute, 1979.

Stufflebeam, D., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond. R. L., Merriman, H. O., and Provus, M. M. Educational Evaluation and Decision Making. Peacock Publishers, Itasca, Illinois, 1971.

Sween, J. A. Experimental Regression Design: Inquiry into the feasibility of nonrandom treatment allocation. Unpublished doctoral dissertation, Northwestern University, 1971.

Wolf, R. M. Selecting Appropriate Statistical Methods. In R. A. Berk (ed.) Educational Evaluation Methodology. Baltimore: John Hopkins University Press, 1981.