DOCUMENT RESUME

ED 102 653 EA 006 770

AUTHOR Calfee, Robert
TITLE The Design of Experiments and the Design of Curriculum. Stanford Evaluation Consortium Occasional Paper.
INSTITUTION Stanford Univ., Calif. Stanford Evaluation Consortium.
PUB DATE Oct 74
NOTE 50p.

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS *Curriculum Design; Curriculum Development; Curriculum Evaluation; Curriculum Planning; *Curriculum Research; *Experiments; Measurement Techniques; *Research Design; *Research Methodology; Statistical Analysis

ABSTRACT

This paper evaluates current practices in curriculum design and discusses some proposals for using efficient Fisherian experimental designs to remedy certain shortcomings in current practices. Two general questions are approached in this paper: (1) how to obtain rational and empirical evidence that a basic curriculum under planning and development will do the job for which it is intended, and (2) how to establish the conditions under which a curriculum can be most effectively "installed" in a particular classroom to meet the needs of a specific teacher and a specific student group. (Author/MLF)

OCCASIONAL PAPERS OF THE STANFORD

Evaluation Consortium

Stanford University, Stanford, California, 94305

# The Design of Experiments
# and
# The Design of Curriculum

Robert Calfee

October, 1974

# THE DESIGN OF EXPERIMENTS AND THE DESIGN OF CURRICULUM[1]

Robert Calfee

Stanford University

A curriculum, by dictionary definition, is a course of study. It can be a collection of books and other materials. It may be a set of teacher manuals, which form the core of many curriculum programs. Curricula can range from scope and sequence charts to detailed writeups, from general discussion of concepts and strategies to exact presciptions.

A curriculum can be viewed as an organized structure for carrying out instruction in some domain. The structure encompasses content (the thing to be taught) as well as time (the order in which things are taught). The temporal structure is especially germane when we think about a curriculum from the student's point of view--a curriculum is something that unfolds over time from one day to the next.

Curriculum construction is a complex operation, entailing numerous decisions. In the case of some programs, the development process is long, involved, and expensive--an elementary reading program entails the work of dozens of people over years at costs in the millions. At the other extreme is the teacher who creates his own program from day to day, using his head and whatever resources are at his disposal. The intermediate case is typical--the teacher uses an established curriculum as a starting point, modifying it as necessary to meet local needs.

This paper addresses the typical case. First the paper will critically review current practices in curriculum design. Then it will

discuss some proposals for using efficient Fisherian experimental designs to remedy certain manifest shortcomings of current practices.

There are two general questions that I hope to answer in this paper. How can we obtain rational and empirical evidence during the planning and development of a basic curriculum about whether it does the job for which it was intended? How do we establish the conditions under which a curriculum can be most effectively "installed" in a particular classroom to meet the needs of a specific teacher and student group?

## Curriculum Development

Planning. At the present time, curriculum programs are created in a relatively unsystematic fashion. There is a planning stage, in which a range of alternatives is considered. A number of curriculum theorists have discussed various ways to approach the task (Tyler, 1949; Taba, 1962; Kirst & Walker, 1971). These approaches vary considerably in clarity, analytic rigor, and practicality. Often the planner simply focuses on one of two specific ideas that he considers innovative and crucial to the success of the program. For instance, he might think that printing vowels in contrastive colors will ensure that beginning readers learn the vowel correspondences of the English language.

Development. In the development phase, a large number of people work together to create the curriculum. Squire (1974) has described this interaction frankly, though perhaps too optimistically:

"How do publishers ensure that the reading materials they publish are usable and workable in the classroom?

"Traditionally they have relied on just about every Research and Development (R&D) resource that has been available to them.

"--They select authors with practical classroom experience and familiarity with classroom applications of research.

"--They engage experienced and successful writers of literature for children, hoping that the writers' demonstrated sensitivity to the interests of children will provide a reservoir of 'insights' useful in writing or choosing selections for reading.

"--They rely on the judgment and insights of professional reading editors, the large majority of whom have devoted their careers to teaching and education, and the staffs in some publishing houses are not too unlike the education faculties in many colleges.

"--They depend in initiating new programs on the accumulated background studies on previously published programs--the elements in programs that worked, the elements that didn't work. It is no accident that the majority of publishers who were strong in reading twenty years ago continue to be strong today.

"--They build on small-scale 'experimental' projects initiated by individual schools and school systems, attempting to make the innovative dimensions of an isolated experiment usable by teachers everywhere.

"--They call on professional scholars and successful teachers to review manuscripts prior to publication, and today especially they call on qualified and sensitive educational leaders to consult on problems of cultural pluralism and sexism in content and graphics.

"--They check the readability level, the concept density, the

interest level, of particular manuscripts prior to publication and they check the authenticity of content.

"--They ask selected groups of children to read and use materials prior to publication to obtain an indication of pupil response.

"--They organize tryouts of especially critical materials prior to publication."

Much happens during development. Everyone involved makes day-to-day decisions that affect the character and effectiveness of the curriculum. Documentation of the growth of the curriculum is sparse and unsystematic. Evaluation--of the quality of the program components, and of the degree to which each component adequately represents the original planning criteria--likewise tends to be informal, and the influence of evaluation on the developing curriculum is a happenstance matter.

Numerous decisions have to be made during the development phase. A theory of instruction (Bruner, 1966, Ch. 3)--and a curriculum can be viewed as a realization of such a theory--should guide certain decisions:

| | |
|---|---|
| Substantive content | What should be taught? |
| The sequencing of content | In what order should things be taught? |
| The method of delivery | What materials and format should be used (books, games, pictures, etc.) |
| Provision for individual differences | How to deal with different entering levels, rates of progress and interests; how flexible should the program be? |
| Assessment | How should learning be measured? What feedback is to be given the student? What criteria will be used to evaluate progress? |

As an aside, it might be interesting to reflect on the relative influence on each of these decisions of the various involved individuals and groups--the author, scholar, publisher, parents, community, board of education, etc.

In principle, the original planning concepts should be preserved during these multitudinous decisions. In actuality, the result often departs in form and substance from the original plan. The new curriculum is then subjected to a series of critical reviews and tryouts, and further changes are made to render it more suitable. This process generally yields a product acceptable in a wide variety of conventional classrooms. If it is not noticeably more effective than other efforts, or if it fails to meet the needs of <u>some</u> teachers and <u>some</u> students--well, no one is perfect.

<u>Evaluation</u>. Just how correct are the decisions in planning and development? And how effective is the final product? Formal evaluation is made after planning and development. To be sure, there is argument and debate, review and critical analysis all along the line. These are often dignified by the term "formative evaluation." All too often this term means that the evidence is weak and the documentation sparse or nonexistent. The reliance on empirical data is generally slightest during the modeling and fashioning of the curriculum, when significant change is still possible. Only after the product is completed and hardened is there any effort to determine effectiveness by actual performance. Summative evaluation, as this latter activity is known, is eclectic in character. The curriculum "as a whole" is evaluated by general measures such as

standardized achievement tests, which may have little relation to the content or purposes that distinguish the new curriculum from its predecessors.

A number of educators have criticized present practices in curriculum evaluation (Scriven, 1967; Cronbach, 1963; Stake, 1966; Wittrock & Wiley, 1970; Bloom, Hastings & Madaus, 1971). Kirst and Walker (1971), summarizing a discussion of current practices in curriculum development, conclude that "curriculum decisions are not based on quantitative decision techniques or even on a great deal of objective data (p. 487)." Walker (1973) ponders the possibility that "many of us in curriculum have at best a comparatively weak commitment to empirical research as a means of dealing with our professional problems (p. 63)." In that paper, he points to self-imposed restraints in curriculum research and evaluation that make much existing work useless--such as exclusively "behavioristic" measures of performance, and the search for isolated, "one-thing-at-a-time" cause-effect relations.

Walker and Schaffarzick (1974) reexamined data from several curriculum evaluation efforts, separating outcome measures that meshed reasonably well with a given curriculum from those that did not. Their general finding was that students do well when tested on the content they have studied, and relatively poorly when tested on content they have not studied--unsurprising but reassuring. Their conclusion is: "What these studies show, apparently, is <u>not</u> that the new curricula are uniformly superior to the old ones, though this may be true, but rather that different curricula are associated with different patterns of achievement (p. 97)." This promising though modest conclusion may represent

the apogee of current research on curriculum.

## The Centrality of Evaluation

Improved evaluation is fundamental to better curriculum development, revision, and installation practices. We do not lack for imaginative, innovative and effective ideas about how to teach; at least this holds for certain basic subject-matter areas like reading and to some extent mathematics. Rather, we lack adequate means for determining which ideas are really good, and which ones are just so-so.

The role of evidence. The collection, interpretation and weighing of evidence should be continuous during the creation of a curriculum program, from planning, through development, to the final stage of installation in classrooms. Wherever decisions have to be made, the basis for a choice can be subjective, political or empirical. If the latter is possible, it should have priority.

Evaluation procedures should be directly linked to pertinent questions at a given stage of development. Expert judgments are useful evidence in many situations; anecdotal classroom observations may be more informative than the quantitative data obtained from standardized achievement tests. But it is important to apply minimum standards to evaluation no matter what context: (a) There should be an empirical basis for the evaluation, and the evidence should be of adequate reliability. (b) The evidence should be documented and capable of substantiation. (c) Evaluation should be based on multiple sources of information.

Analysis of a problem. Next, consider performance-based evaluation, in which behavioral data from students or teachers is collected as part of evaluation. In investigating any complex system, scientific progress

often depends upon analysis of a complex problem by dividing it into subproblems that can be studied independently. We have relatively few guidelines as to what partitionings are suitable in curriculum research (Walker, 1973, p. 68). For instance, the best way to teach a child to solve quadratic equations may depend on whether he learned to add by rote flash-card drill or by counting on his fingers, but this seems farfetched. A priori judgment may have to guide us in deciding what components are and are not independent until we have a more adequate empirical base than at present. Elsewhere Floyd and I have discussed the use of multifactor experimental designs to establish the independence of various elements of a curriculum at the same time that evaluative data are being collected (Calfee & Floyd, 1972).

Experimental control in planning. It seems vital to progress in curriculum evaluation that experimental control be established over the major decision factors in a curriculum plan, and over subsidiary factors where feasible. Most often a new curriculum represents a single set of decisions about content, sequence, method of delivery, individualization and assessment. If a new curriculum incorporates a fixed set of decisions for each component, we have no way of obtaining evidence about the outcome under alternative decisions. A comparison of two curricula in which choices are varied in an unsystematic manner is also uninformative, because of uncontrolled confoundings.

To see where and how experimental curriculum research might be done, let us consider the process of curriculum development. The initial phase involves a structural description of the curriculum to identify and label the decisions actually embodied in the curriculum--decisions as to

learning goals, organization of content, instructional sequence, and necessary entry behaviors for beginning any sequence (Figure 1). This should provide a clear picture of the decision structure underlying the curricular sequence. In this phase, the designer thinks about the assumptions behind particular teaching methods, content, and materials. He assigns priorities to various learning goals.

Once certain critical decisions have been identified, alternatives that are feasible and worthwhile can be proposed. The third phase centers on the design of parallel experimental curriculum strands (Figure 2). Parallel strands are built around elaboration of alternative pathways at critical decision points, so that experimental variation is introduced at loci judged to be of potential significance. This analysis requires efficient design for control of multiple factors.

Evaluation in the classroom. When a curriculum is tried out under real classroom conditions, it is important to control external factors, such as variations in the teacher, the students, and the school environment. An important question in the evaluation of any new curriculum is the degree to which it is effective under the varied conditions that arise in real classrooms. Control over external variation requires that the researcher identify potentially relevant factors, and that he select a sample of schools, teachers, and students in which these factors are represented in a design that allows the isolation of effects associated with such factors. By incorporating control over external factors in the evaluation design, it is possible to measure specific interactions between curriculum factors and external factors (this is, in a slightly different guise, what is called aptitude-treatment interaction), and

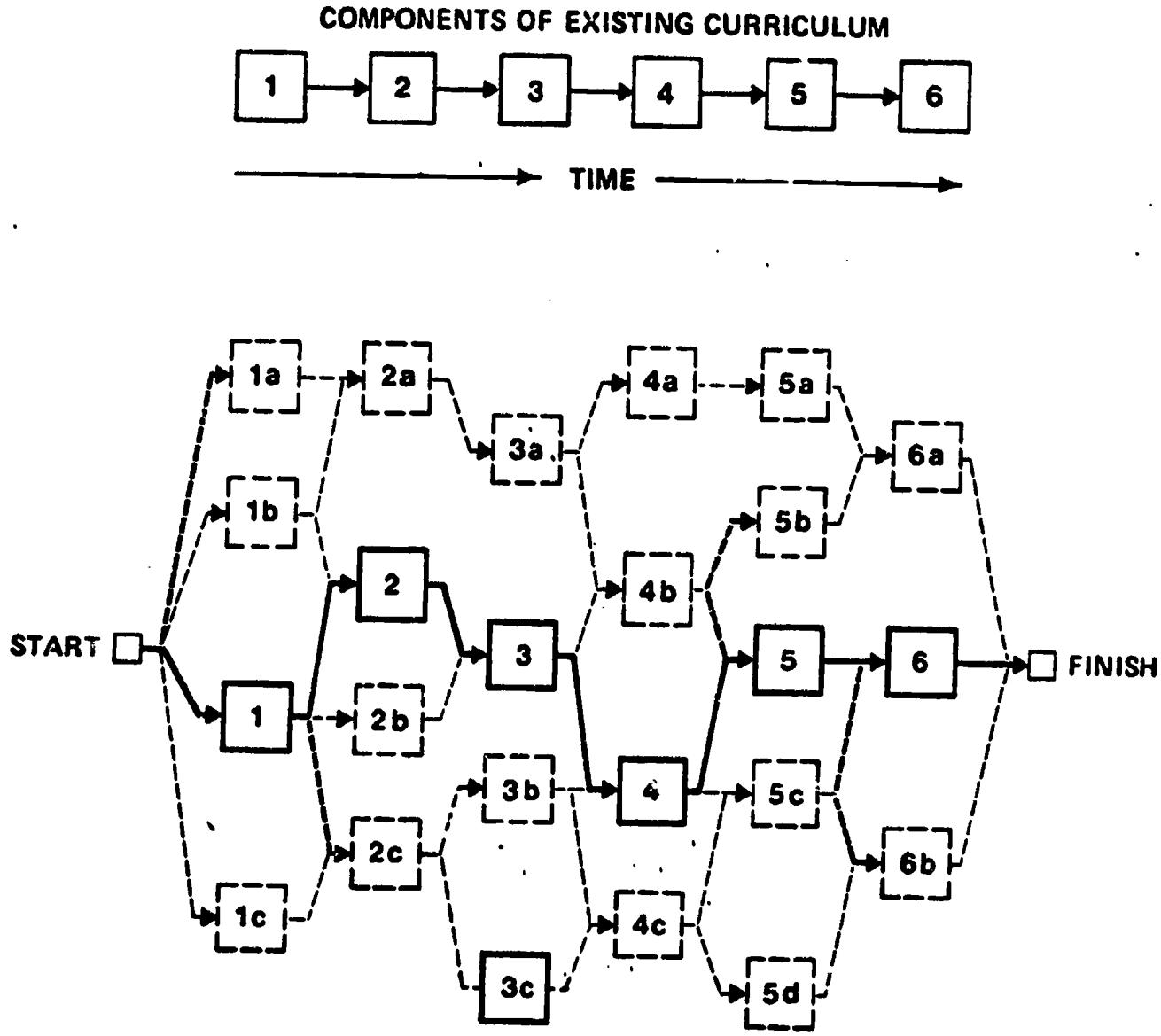


FIGURE 1 - Structural Analysis: At the top are blocks representing the sequence of components (or lessons) of the existing curriculum. One block might represent a module in which the child is taught a set of "sight" words, or caught the principle of adding two-digit numbers with a carry. Below, these components are placed in a decision network illustrating a hypothetical set of choices made by the curriculum planner. Thus, 1a, 1b, and 1c along with 1 were all candidates for the first component. In choosing 1, the planner made a decision which limited the available components for the next stage to 2, 2b, and 2c. The most significant decisions, as identified by the designer and other experts, are indicated by dotted lines. These serve as the basis for constructing alternative curriculum strands for experimental evaluation.

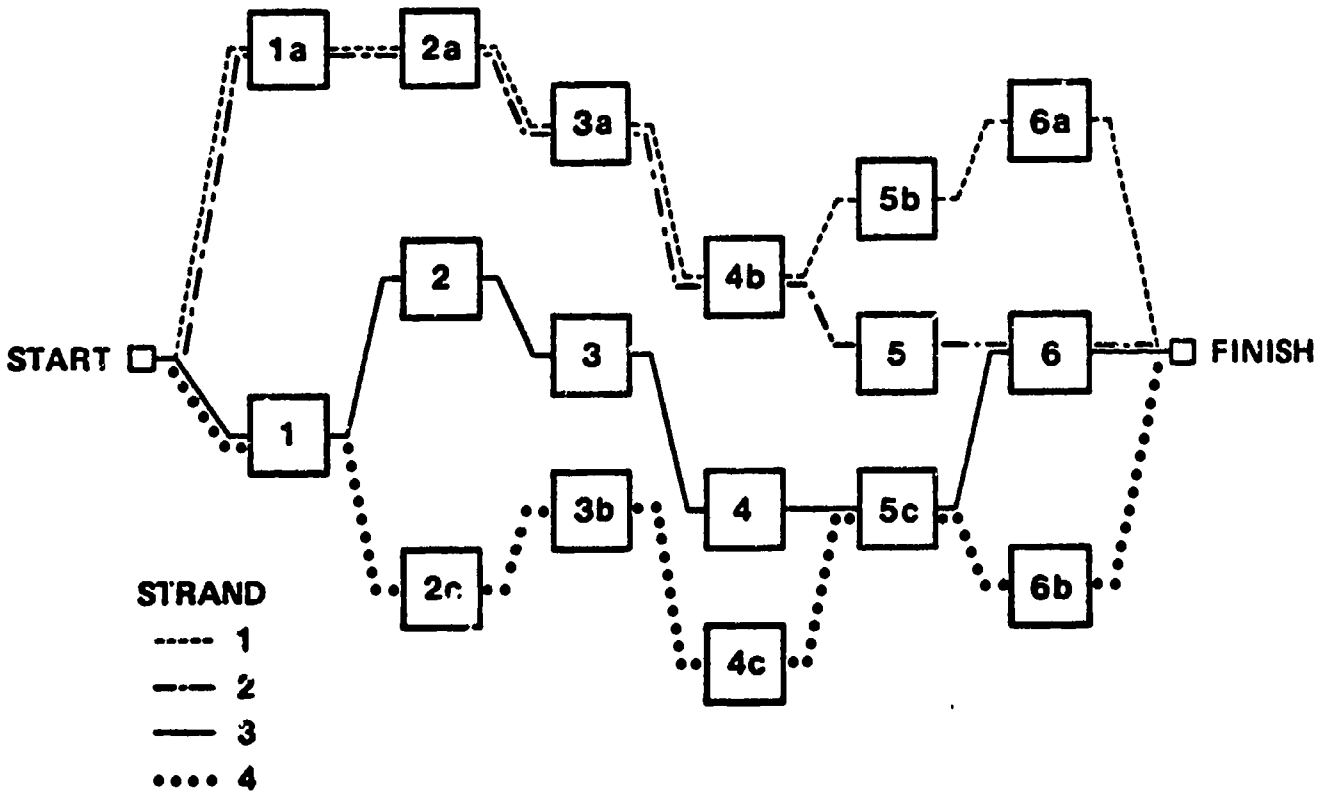


FIGURE 2 - Design of Alternative Curriculum Strands: Four strands are designated in this example. If we assume that mastery has been achieved at each stage, then all students at a given node can be independently assigned to the nodes leading to the next level. This possibility prevents the design from increasing exponentially. Major interactions thought to be important can be investigated, and others are eliminated by default. Ideally, the sequential patterns are less intricate than in this example, so that the modules at each stage would form an independent experimental design structure.

to estimate the magnitude of generalized interactions (for example, curriculum-by-school variability). If the specific interactions are large, this calls for alternate versions of the curriculum, and specification of the conditions under which one or another version is most effective. If generalized interactions are large, then the "transportability" of the decision is altogether questionable.

The First Grade Cooperative Reading Study (Bond & Dykstra, 1967) exemplifies the problem. Six different reading programs were compared, each tested by several project teams in a number of schools. Variability between schools within a project, and variability within projects within a program, were both as substantial as between-program variability. In short, the program distinctions were not significantly related to performance outcomes.

Once more, the establishment of suitable evaluation conditions rests on the adequacy of the experimental design.

Measurement and evaluation. Next, there is the task of constructing an appropriate measurement system. I will focus my remarks on student performance measures, but the same considerations apply to teacher measures, classroom measures, and any other source of information about curriculum effectiveness (e.g., judgments from expert observers, including curriculum specialists, anthropologists, etc.)

A measurement system should rest on an analysis of the essential component processes or elements in learning. In social studies, for instance, suppose that we were to identify as significant learning components (a) a method for selecting relevant historical facts from a passage, (b) techniques for organizing and memorizing such facts, (c) a

body of knowledge (historical facts) of special importance, (d) procedures for critical analysis of a source of historical information, and (e) techniques for comparing and contrasting two sets of historical data. These are not the only important elements in a social studies curriculum, but they are a reasonable starting point. These are cognitive skills--not behavioral objectives. They are ways of thinking and solving problems; they are knowledge. They subsume sets of specific behavioral objectives.

Of potential relevance to the analysis of component skills is current research on information processing (Sternberg, 1969; Anderson, 1970; Kavanaugh & Mattingly, 1972; Lindsay & Norman, 1972; Chase, 1973; Haber & Hershenson, 1973, Ch. 7). Information-processing models take the form of sequentially or hierarchically organized structures of cognitive processes. After postulating a structural model for a given task, the psychologist identifies the specific processes and factors affecting each, and then formulates experiments to obtain evidence about the functional independence and operation of these processes. Sternberg proposed a simple and elegant paradigm for tackling this problem in which a factorial design is built around various combinations of within- and between-stage factors. If the independent-process analysis is correct, performance in a given stage will depend only on variation in factors associated with that stage; factors associated with other stages will not affect the performance of this stage either directly or by way of interactions. (For an extension of this technique, cf. Calfee, 1970, 1974; for a different approach to the same problem, see Carroll, 1974.)

In the social studies example, this approach would require that we

answer two questions for each of the five components: What experimental variables are likely to influence this component directly? How can the operation of the component be measured? For instance, in component (b), techniques for organizing and memorizing facts, a relevant factor might be the method by which a student is taught to organize and memorize. One method might be to arrange the facts in a hierarchical structure and teach by rote repetition; another might be to organize the information in any way and use mnemonic techniques such as the method of loci for memorization. Measurements of this process might include asking the student what he was doing, examining the organizational character of the protocols, or measuring total recall of a body of historical facts, either immediately or after a delay. If memory is a process in the acquisition of social studies, a factorial design including a range of variables should reveal that recall is affected only by memory factors, and not by factors affecting other components.

From the perspective of the cognitive psychologist, assessment batteries created in this fashion are factorial experiments. From the perspective of the educational psychologist or curriculum evaluator, these batteries can be viewed as tests or assessment instruments. The data can be examined in a straightforward way to answer questions about the relative importance of each factor, and about sources of substantial individual differences. There is no need to resort to factor analysis, or attempt to construct "factor-pure" tests. The experimental design is self-confirming and self-correcting with regard to the validity of the underlying process model.

Fisherian Experimental Designs in Curriculum Planning and Evaluation

At each of the points touched on above, one requirement for establishment of adequate control was the application of experimental design procedures. The past three decades have seen the widespread acceptance of factorial designs in psychology and education. Hierarchical designs are now commonplace, and Latin and Graeco-Latin squares are in frequent use for control of nuisance factors. There is increasing sophistication in the statistical and interpretive analysis of such designs, especially with regard to questions of generalization to various populations (Cronbach, Gleser, Nanda & Rajaratnam, 1972).

The approach has its detractors (cf. Stufflebeam, 1971, for a review of this issue). Some argue that experimental control over school-related research is impractical or unnecessary, and that those factors over which control can be maintained are likely to be trivial. Others confuse design with analysis, and promote multiple regression as preferable to analysis of variance techniques. Sometimes one procedure will do a better job, sometimes the other, but they are based on the same underlying model, and used properly both techniques ordinarily give a similar answer.

The following points about Fisherian designs and analysis of variance bear specifically on the evaluation of curriculum and instructional programs:

(a) The model for Fisherian designs, the general linear model, provides a simple and elegant model where theory is vague, misleading or altogether lacking.

(b) The a priori arrangement of factors into orthogonal structures

substantially increases the sensitivity of a curriculum experiment to questions of interest, compared to naturalistic research. With *a priori* control, factors are likely to be partly or fully confounded, and while techniques exist for *a posteriori* "adjustment" of data, none of these has the power of a balanced, orthogonal design.

(c) Fractional designs, a natural extension of full factorial designs, provide all the advantages of orthogonality, but permit a relatively small amount of data to be used to answer a large number of questions.

The general linear model. In its most general form, the linear model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \varepsilon$$

where Y is a criterion measure, the $X_i$ are factors to be used in predicting Y, $\beta_i$ is the weight of factor $i$, and $\beta_0$ and $\varepsilon$ are baseline and error parameters, respectively (Morrison, 1967; Rao, 1965; Scheffé, 1959). The model can be formulated for the multivariate as well as the univariate case. It is the basis for analysis of variance and multiple regression analysis, both of which are flexible and robust techniques.

The power of the linear model as a substantive model is often overlooked (Suppes, 1974). Current applications emphasize tests of the null hypothesis, but the machinery exists for more explicit tests of parameter values and for measuring the relative influence of factors in a set by examining components of variance.

The linear model also provides a readymade system for handling several other problems: What measurement scale provides the most parsimonious description of a set of data? What is the magnitude of

interactions among a set of basic predictor factors? What is the magnitude of confoundings among predictor factors?

In short, the linear model stands ready to serve the needs of curriculum research. It is able to handle the complexities of this area in a flexible and statistically powerful fashion. Theory and techniques of analysis have been worked out in great detail.

A priori design and a posteriori analysis. Fisherian designs require strong control over the selection of experimental units and the assignment of treatments to units. These requirements weigh heavily on the shoulders of those of us who try to do educational research. Schools are distrustful and cynical about the value of research, and protective of themselves. It is easier to find some school that will let you carry out research than to try to gain entry to the school called for by the design. It is easier if a teacher volunteers for a treatment than to arrange for a teacher to follow a treatment according to a design. The path of least resistance for the investigator is to simply look around for preexisting treatment-unit combinations, and then to see what design he has.

What is wrong with the path of least resistance? The answer lies in the presence of substantially confounded factors, with consequent lack of control. Not everyone agrees that a priori design control is important. Cohen (1970) has proposed techniques for "naturalistic" design. A posteriori adjustment of confounded data is possible under certain highly restrictive assumptions (Elashoff, 1969). And the possibility of causal inference from correlational data has been considered (Wittrock and Wiley, 1970). However, a posteriori techniques are much weaker statistically than comparable a priori designs. They rest on

strong assumptions, frequently untenable. And they are of little use to the curriculum planner and developer, because the natural process of curriculum development provides little "natural" variation of the sort needed for empirical testing of specific hypotheses.

## Fractional Designs and Efficient Multifactor Experiments

Educational and experimental psychologists usually think of a factorial design as a between-subjects arrangement consisting of all combinations of two or three factors, with 10 or more subjects per combination. "As everyone knows," it is essential to have a large number of subjects per combination, because if the sample size is too small then the statistical test will be weakened. In fact, this "sample-size-per-cell" requirement is based on misunderstandings about what is being tested in factorial designs.

The high cost of this paradigm severely restricts the experimenter's ability to investigate complex problems. With three two-level factors, there are 2 x 2 x 2 or eight combinations, which at 10 subjects per combination amounts to 80 subjects. The increase is geometric with additional factors and levels per factor. Designs with more than five factors are impractical because of the cost--even if the subjects are students. If they are teachers or programs, the constraints are even tighter, and designs with one or two factors are about the limit.

A contrasting paradigm is the within-subject design, now common in behavioral experiments. When each subject is tested under several factorial combinations the statistical tests are quite sensitive. Usually the subject is tested only once under each combination, and the "sample size" requirement is conveniently ignored. Control for order and

materials variables is usually achieved with a Latin square or Graeco-Latin square. The analysis of correlated data from such designs raises problems, but techniques for handling these are being continually refined. All in all, within-unit designs constitute a powerful methodology for studying certain behavioral questions.

Even here, there are practical limits to the number of factors that can be examined, as long as full factorial designs are employed. A design with six two-level factors has 64 combinations. If each treatment combination takes a minute or more, the "50-minute hour" limit typical of much psychological research is violated. And in curriculum research, we may be talking about treatment combinations that require days, weeks or months to administer.

Fractional designs provide an efficient alternative to full factorial design. These designs are not new; the basic procedures have been available for at least 40 years, and are described in a number of standard texts (Kirk, 1969; Winer, 1971). For some reason, they have seen little use in the behavioral sciences, except for the special case of Latin Squares.

Two related design procedures comprise fractional designs: fractional-factorial and confounded-blocks designs. In a fractional-factorial design, the experimenter selects a balanced fraction of cells from the full design. In a confounded-blocks design the full design is divided into orthogonal blocks, each of which is assigned to a different experimental unit. Many applications involve combining these two design techniques; a fraction of the full design is selected and then broken into blocks, each of which is assigned to a different unit.

The basic concepts of fractional-factorial and confounded-blocks designs with two-level factors will be illustrated by an example in which there are three two-level factors, as shown in Figure 3A. In a $2^3$ design, eight degrees of freedom are available for estimating the grand mean, main effects, and interactions, each with one degree of freedom.

The two sets of cells in the design in Figure 3A labeled + and - represent the two halves of the full design defined by the ABC interaction. The ABC interaction, which in this example has been used to divide the full $2^3$ design into two balanced chunks, is called the defining contrast. Consider the consequences of carrying out an experiment using only the + cells of the design. There is one degree of freedom for estimating the grand mean, and three degrees of freedom for estimating treatment effects. The ABC effect in this fractional design is the same as the grand mean; hence, information on ABC is lost. Furthermore, the estimates of the following pairs of effects are also identical and hence confounded:

$$A = BC$$

$$B = AC$$

$$C = AB$$

Two confounded effects such as A and BC are referred to as <u>aliased</u>. Figure 3B shows the analysis of variance source table for this design. Each source is redefined in terms of the aliasing patterns, using ABC as the defining contrast. The cost of cutting the full design in half is that information about interactions is lost, and so the experimenter must think seriously about what hypotheses are really worthy of

FIGURE 3A - A full $2^3$ design, with three treatment factors A, B, and C each at two levels, 0 and 1. Cells containing +'s and -'s represent the two levels of the ABC interaction.

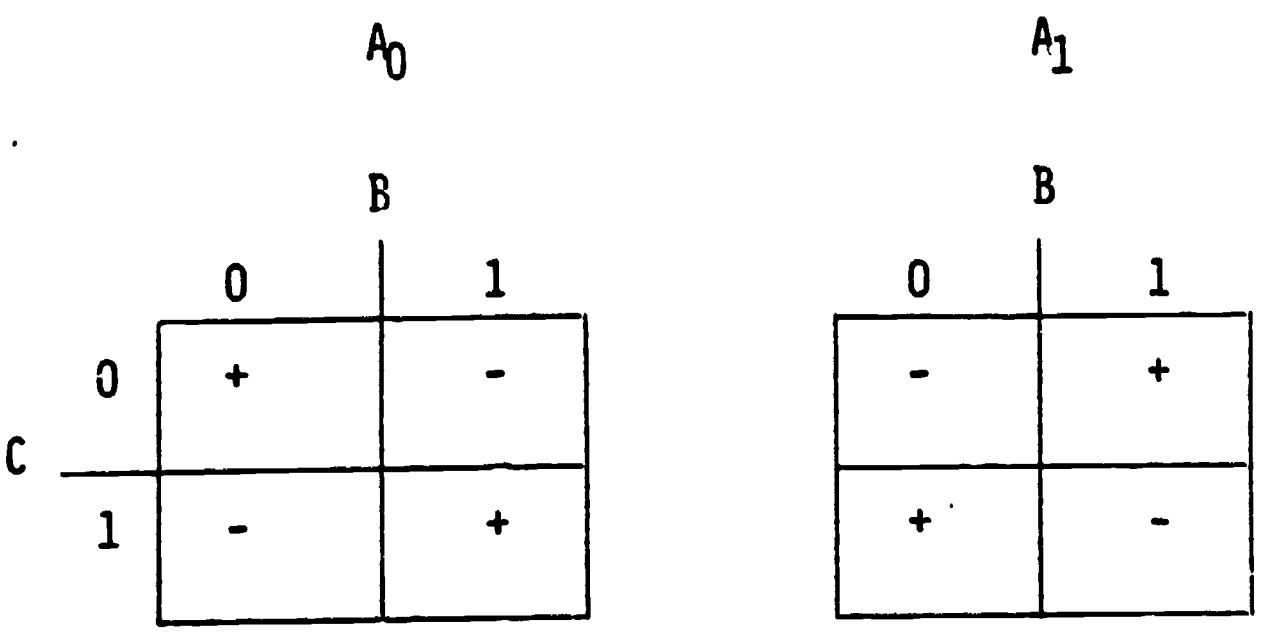


FIGURE 3B - ANOVA source table for a ½-replicate of a $2^3$ design, with ABC as the defining contrast.

| SOURCE | DF | ALIAS |
|---|---|---|
| MEAN | 1 | ABC |
| A | 1 | BC |
| B | 1 | AC |
| C | 1 | AB |

FIGURE 3C - ANOVA source table for the analysis of a $2^3$ design run in two blocks. Each level of the blocks factor, X, was assigned to a different experimental unit, and is therefore a between units factor.

| SOURCE | DF | ALIAS |
|---|---|---|
| MEAN | 1 | |
| A | 1 | BCX |
| B | 1 | ACX |
| C | 1 | ABX |
| AB | 1 | CX |
| AC | 1 | BX |
| BC | 1 | AX |
| ABC | 1 | X |

investigation.

Next let us look at a confounded-blocks design based on the same $2^3$ design. In this case, the ABC interaction is used to split the full design into two fractions or blocks, the +'s and -'s of Figure 3A, each of which is assigned to a different experimental unit or subject. There is one degree of freedom for the grand mean, and seven degrees of freedom available for estimating treatment effects. We have in effect created a new dummy variable for blocks, designated by X, each level of which is associated with one level of the ABC interaction; $X_0$ is assigned to the +'s in Figure 3 and $X_1$ to the -'s. Since each level of this variable has been assigned to a different experimental unit, it constitutes a between-units effect. The aliasing patterns for the remaining factors, all within-units effects, are:

| | |
|---|---|
| A = BCX | AB = CX |
| B = ACX | AC = BX |
| C = ABX | BC = AX |

The analysis of variance for the confounded-blocks design is shown in Figure 3C.

## A Curriculum Experiment in Beginning Reading

Our first example looks at the process of planning a curriculum experiment in beginning reading. Suppose that the curriculum can be represented in modular form. Many reading curricula are now constructed in this fashion, in the sense that within each lesson there are sub-sections dealing with specific tasks (Figure 4). We want to focus on three curriculum components: content, materials and format, and management system. The curriculum is designed around a set of texts, the

target population is first graders, and the curriculum is supposed to meet the needs of a variety of teachers and students.

A preliminary revision of an existing curriculum has been planned, and the purpose of the experiment is to provide information about the merits of various elements in the revision. There are 16 two-week segments in the curriculum, and the curriculum developer thinks it feasible to create as many as 16 variations on the basic curriculum program, and to carry out the experiment in 32 classrooms.

Preliminary discussions have identified the following major questions:

(A) Content decisions: Basic reading instruction

a) What is the value of a relatively strong emphasis on phonics/decoding skills, versus a relatively strong emphasis on comprehension/"reading for meaning?" The planner's intention is to incorporate both components in the curriculum, but he would like some information on the degree to which teachers make use of the two types of materials, and the amount of learning and student acceptance of these two types of materials at different times in the school year.

b) Within the two levels of the preceding question two subquestions are nested:

(1) Is phonics most effectively presented by a rule orientation based on learning letter-sound associations and blending procedures, or by a word-based orientation à la Bloomfield and Barnhart (1961)?

(2) How important is vocabulary control? Is reading for

meaning better taught with high-frequency words that are likely to be irregularly spelled versus less frequent words that are regularly spelled?

(B) Content: Skill development

a) Does it make a difference whether or not visual-skills work sheets and other similar materials are included in a module?

b) Ditto for auditory-skills materials?

(C) Content: Literature

a) Does it matter whether or not materials for story-telling and poetry are included in a module?

b) Ditto for creative dramatics, writing, etc.?

(D) Materials/Format

a) Does it make a difference whether or not student work-books are included in addition to the basic textbook materials?

b) Ditto records, audio tapes, films, etc.?

c) Ditto supplementary materials especially designed for very fast and/or very slow learners?

(E) Management

a) Does it matter whether a module is constructed around a learning-to-mastery emphasis, as opposed to a minimal-competence or remedial model?

b) Does it matter whether or not an assessment system is provided?

c) Ditto a record-keeping system?

The questions in (A) are of fundamental importance to the construction and refinement of curricular content in the final version of the program. The questions in (B) through (E) are all yes-no questions. These relate to the tendency to throw everything, including the kitchen sink, into current curricula, a smorgasbord approach. This costs money, makes it more difficult for a teacher to identify useful components, and is of uncertain benefit. The answers to (B) through (D) will provide evidence on which auxiliary components are worthwhile additions to the basic curriculum.

The preceding questions comprise a set of twelve two-level factors. To create a particular instructional module, we would have to consider twelve decisions, each of which might be made in either of two ways. For instance, one module might have (A.a) a strong phonics emphasis, (A.b.1) a rule-orientation, (A.b.2) no visual-skill materials, but (B.a) auditory-skill materials, (B.b) story-telling materials, and (C.a) creative dramatics materials, (C.b) no workbooks, (D.a) no audio tapes and (D.b) no supplementary materials for fast/slow students, but (E.a) a learning-to-mastery emphasis with both (E.b) an assessment system and (E.c) a record-keeping system.

Besides the twelve planning or treatment factors described above, it is important to control order, the time in the school year when a given module is presented. Assume that the school year is divided into four chunks, and that the order of module presentation is balanced within and across chunks. Assume further that each child goes through 16 ($= 2^4$) modules during the school year, and so the full design comprises $2^{16}$ combinations. Sixteen balanced versions or blocks of the

basic curriculum are to be constructed, and so a $2^8$ fraction of the full design is required.

$2^{12}$ (Treatment Factors) x $2^4$ (Order) = $2^{16}$ (Full Design) = $2^4$ (Segments) x $2^4$ (Blocks) x $2^8$ (Fraction)

This design can be planned in such a way that within-class tests of each of the main effects are possible, as well as selected interactions. (In fact, only seven of the 120 two-way interactions are not measurable.) For instance, it might be useful to ask about the effectiveness of phonics versus meaning early in the school year compared to later in the year. The relation of assessment and record-keeping materials to mastery learning poses another interesting interaction question. The point is that one can handle this complex of problems in a relatively sensitive design, with adequate control over the entire set of factors, including order, at a cost that is feasible.

It was assumed earlier that a set of 16 variant curriculum programs was to be installed in 32 classrooms. Each classroom will have students who vary in entering ability level, sex, and other pertinent factors. It would make sense to use such student information in the analysis. This would permit an especially strong attack on what has been referred to as the aptitude-treatment interaction hypothesis (Cronbach & Snow, 1969).

An even more interesting possibility presents itself. Suppose we wanted to find out how the effect of curriculum decision depends on preexisting characteristics of the school and teacher. We would need to plan a between-class design that provided control over

relevant factors that differentiate schools and teachers. For example,

(A) School characteristics

(a) Urban/suburban

(b) High/low socio-economic neighborhood

(c) Self-contained/open-school plan

(B) Teacher characteristics

(a) Experienced/beginning

(b) Prefers to follow curriculum and teacher manual closely/ prefers to adapt curriculum to own program.

(c) Prefers large group instruction/small-group, independent work.

These six two-level factors, together with the four block factors, constitute a $2^{10}$ design. By planning a $2^5$ fraction for the 32 teachers available, control is maintained over school and teacher factors and the assignment of curriculum factors to classes. The main effects of school and teacher factors are all testable. Equally importantly, it is possible to test hypotheses about interactions between school-teacher factors and curriculum factors. For instance, what is the effect of a learning-to-mastery component for teachers who prefer large group instruction, compared to those who adopt a more individualistic approach?

The major point here is that it is feasible to plan designs that handle the complexities that arise in curriculum planning and development, and that achieve the rigor of control deemed necessary in behavioral experiments.

The experiment described above constitutes a broad framework within which another level of experimental questions could be planned. For

instance, within a phonics module one could raise questions about (a) the order in which specific letter-sound correspondences are presented, (b) the rate at which new correspondences are introduced, or (c) maintenance of constancy or variability in vowel patterns (e.g., are short vowel patterns presented first followed by long vowels, or are both long and short vowels presented contrastively in a single session?). By building successively more detailed designs, the experimenter can create a hierarchy of experiments, within which might be embedded experiments as precise as those now conducted in experimental psychology laboratories. In the larger context of the entire study, such studies might achieve a degree of relevance and generalizability that they now lack.

An Experimental Study of a Curriculum for Increasing Teacher Effectiven

As a second example, consider an experiment designed as a part of study of teaching at the elementary level.[2] The "subjects" are classroom teachers. The curriculum is a series of modular units, each of which focuses on a single teaching skill area. There is special intere in the most efficient means of "delivering the message." Training must relatively fast, and acceptable to the majority of the teachers.

The research design to be presented can be thought of as a combination of experimental and case-study methodology. Each teacher is to be studied over a full school year. At intervals, the teacher is trained on a specific instructional skill. Classroom observation provides the major data on the effects of each training procedure. Because of the intensive nature of training and observation, only a small number of teachers can be studied. The design to be presented is intended to be

illustrative; alternatives to these particular design factors could be (and would be) given serious consideration.

The primary question is: What is the effect on classroom teaching of short-term, intensive training of teachers on specific classroom practices? For example, suppose that more effective teachers provide differentiated, task-specific feedback to students, and also carry out continuous assessment of student progress. If teachers are given training on each of these skills, which ones provide immediate payoff as measured by a noticeable change in classroom practice?

Second, what is the relative effectiveness of different methods of training teachers in effective classroom teaching practices? For any skill in which training is needed, several approaches can be used--traditional inservice methods, demonstration classes, audiovisual and television equipment, among others. Variation in the "delivery system" will provide a test of the relative effectiveness of different training procedures.

Let us assume that a maximum of sixteen teachers can be studied, and that four training modules are to be administered to each teacher during the school year. There are two basic steps in preparing the experimental design: First, deciding what factors to use in selecting the sample of teachers and schools, and, second, deciding what factors are important to the substantive content of the training modules and the method of delivery. These will be referred to as the <u>between-teacher</u> and <u>within-teacher</u> plans, respectively.

Here are a set of illustrative between-teacher factors for this study:

School Factors

A. Socioeconomic status of neighborhood served by school

0 above median

1 below median

B. Area served by school

0 urban, high density

1 suburban, small town, rural

C. Administrative climate and control

0 high control by principal of instructional program

1 low control by principal of instructional program (laissez faire)

Teacher Factors

D. Grade

0 Primary

1 Elementary

E. Teaching Style

0 relatively structured instructional practices

1 relatively unstructured instructional practices

F. Student outcomes in previous years

0 negligible difference between actual and predicted gain

1 large positive difference between actual and predicted gain

Other factors might be considered as serious candidates; the problem is at least this complicated, and maybe more so. Let us take this as a starting point, and suppose that our task is to plan a study with the six factors above for a group of sixteen teachers. The full design calls for $2^6$ or 64 teachers. We can include 16 in the design, and so a

$2^2$ or ¼ fraction must be selected from the full design.

Figure 4 shows a plan for a one-quarter replicate of a $2^6$ design based on the school and teacher factors described earlier. We start with the full $2^6$ or 64-cell design. Two high-order interactions, ABCD and CDEF, are used to divide the full design into four balanced sets. In the upper half of each cell is a + or - indicating whether that particular cell is positive or negative in the ABCD interaction. In the lower half of each cell is a + or - for the two halves of the CDEF interaction. Each of these interactions divides the full design into two halves, both of which are balanced with respect to each other. The two interactions taken together divide the full design into four pieces, all four pieces balanced with regard to each other.

Each quarter is represented by a + or - for ABCD, and a + or - for CDEF. Note in the figure that there are exactly 16 instances of each of the four patterns, (++), (+−), (-+) and (--). One of these quarters, the one with (--) in each cell, was selected at random for this experiment. Any one of the four quarters would do equally well. These sixteen cells are outlined in the figure. You can see the symmetry from one quadrant to the next which reflects the balancing.

The design requires two teachers in each of eight schools. For instance, a low income, urban, high administrative control school will be selected in which a primary teacher is teaching in a relatively structured fashion, with above average gain in student performance; an elementary teacher will also be selected who uses a relatively unstructured approach to instruction, and whose students are also relatively higher in performance than predicted.

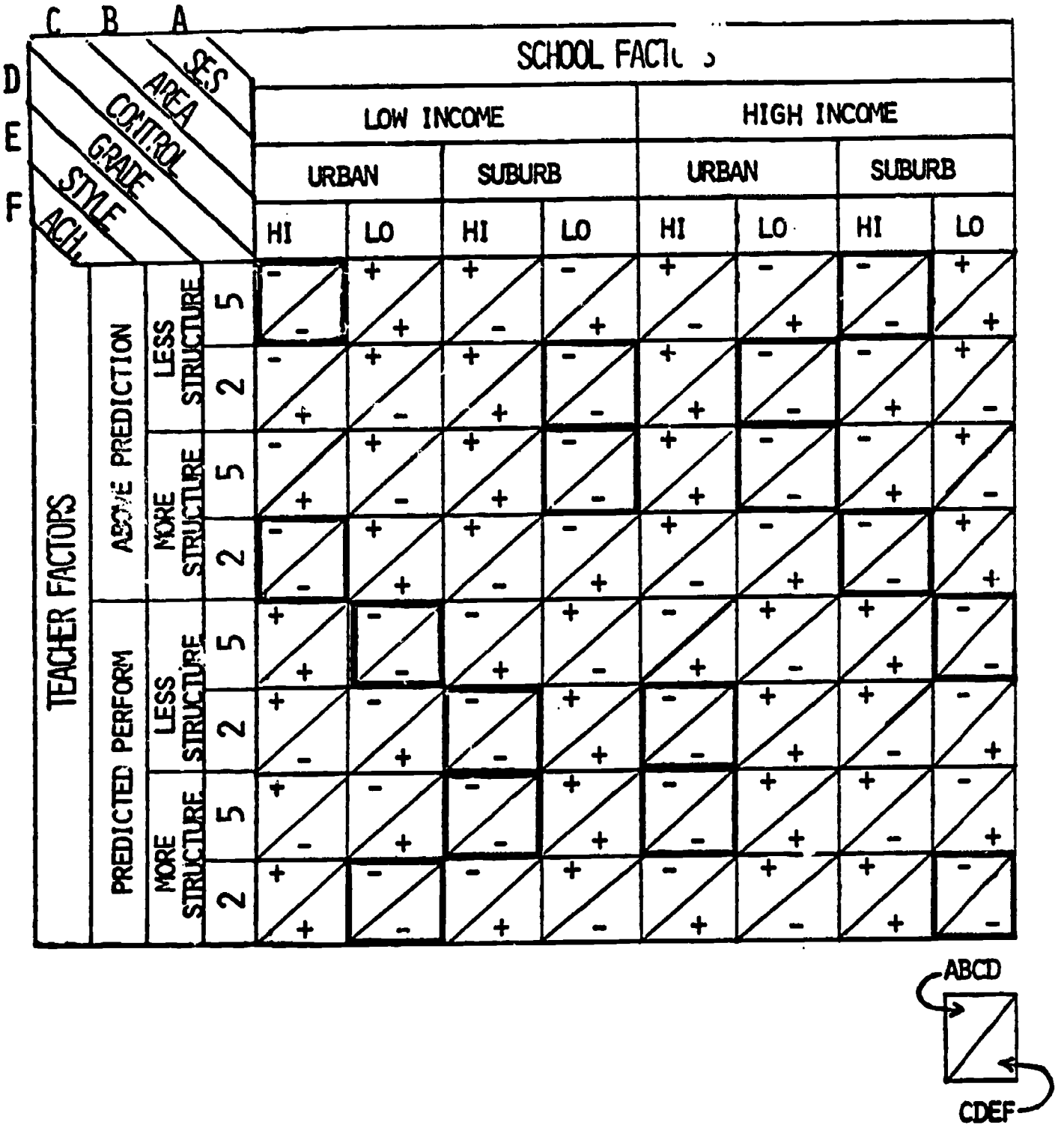


Figure 4

Plan of $2^6$ experiment, showing how ABCD and CDEF interaction devide 64 cells into quarters. the $^{-}/_{-}$ quarter is shown in bold as a suitable fraction for conducting an experiment.

In this study, we are mainly interested in the main effects of the school and teacher variables, and except for a few two-way sources, interactions can be disregarded. The primary purpose of this experiment is not to examine differences between teachers. We need control over factors associated with differences between teachers, and would like to know the relative magnitude of the main effects of these factors. But the chief purpose of the study is to examine the effectiveness of training programs, and to see whether or not training programs are differentially effective as a function of teacher variables. This is accomplished by the within-teacher portion of the design discussed later.

Here are the sources in the analysis of variance that can be tested with this design, and the hypotheses corresponding to each source:

| Source | | Question: Is the effectiveness of the training program different for... |
|---|---|---|
| A | Socioeconomic status | teachers working in above/below median income schools? |
| B | Area | teachers in urban/suburban schools? |
| C | Administrative control | teachers in a school in which principals exert considerable influence on the instructional programs/schools in which principals exert little control? |
| D | Grade | primary/elementary teachers? |
| E | Teaching style | teachers who employ more/less structure in instruction? |

| | | |
|---|---|---|
| F | Student outcome | teachers whose students have done relatively well/teachers whose students have done about as predicted? |
| AB | | teacher in urban/suburban schools, depending on whether neighborhood socioeconomic level is above/below median? |
| CE | | teachers in "high"/"low-control" schools depending on their teaching style? |
| BF | | teachers in urban/suburban schools, depending on whether the relative student performance is above predicted/ about as predicted? |

The interactions selected for hypothesis testing in this design are for illustration only. As many as two or three interactions could be selected for testing, and there would still be seven or eight degrees of freedom for an error variance term, sufficient to estimate the error variance for this portion of the design.

In the within-teacher portion of the design, three different sets of factors are being proposed for variation across the four training modules for each teacher. These include the area of training, the method of training, and the time when training is administered.

Aera of Training

(G,H) 0 specific, differentiated feedback

1 use of performance-based evidence (as opposed to opinion) in assessment

2 procedures for continuous monitoring in reading and mathematics

3 how to keep records and use them for individualization

Methods of Training

(J,K) 0 traditional in-service training

1 demonstration classrooms

2 television demonstration, microteaching, etc.

3 use of school resource personnel to install and support the program

Time of the Year

(L,M) 0 October

1 January

2 Mid-February

3 Early April

The four-level factors are each represented by two two-level factors for convenience in planning the experimental design.

In the within-teacher portion of the design, we must solve the probelm of fitting a $2^6 = 64$ cell design to the constraints that each of the 16 teachers receives four training modules, each of which comprises one cell of the design. The task of planning is sufficiently complex that plans provided by the National Bureau of Standards (1957) were used. These plans provide the details of how to organize fractional-factorial and confounded-blocks experiments with as few as five factors and as many as 16, for a wide range of fractional and confounded blocks

constraints.

Eight blocks of four cells each are to be selected in a balanced fashion from the full $2^6$ or 64 cell design. This is done by first dividing the design into two halves, as shown in Figure 5, using the 6-way interaction as a contrast. One of the halves, the (-) portion in this example, is then divided into eight blocks of four cells each in a balanced fashion. A block comprises a balanced ordering of four combinations of area of training, method of training, and time of training, which can be assigned to two teachers in one of the eight schools. The block numbers are shown in the figure. In Figure 6, the design has been written out in a different way to show the sequence and combinations of training conditions in each of the eight blocks.

Each of the eight blocks is assigned to one of the schools, and so each pair of teachers goes through a unique training sequence. The design allows all of the main hypotheses of interest to be tested. These are shown below:

Within-Teacher Analysis

| Source | Question: What is the effect on teacher practices of training |
|---|---|
| G, H Area (df = 3) | . . . in different areas such as how to use differentiated feedback, continuous monitoring, etc. |
| J, K Method (df = 3) | . . . using different methods of |

| | Feedback | | | | Evidence | | | |
|---|---|---|---|---|---|---|---|---|
| | Trad. | Demo. | T.V. | Resrc. | Trad. | Demo. | T.V. | Resrc. |
| Oct. | - 1 | + | + | - 6 | + | - 4 | - 7 | + |
| Jan. | + | - 7 | - 4 | + | - 6 | + | + | - 1 |
| Feb. | + | - 2 | - 5 | + | - 3 | + | + | - 8 |
| Apr. | - 8 | + | + | - 3 | + | - 5 | - 2 | + |
| Oct. | + | - 3 | - 8 | + | - 2 | + | + | - 5 |
| Jan. | - 5 | + | + | - 2 | + | - 8 | - 3 | + |
| Feb. | - 4 | + | + | - 7 | + | - 1 | - 6 | + |
| Apr. | + | - 6 | - 1 | + | - 7 | + | + | - 4 |
| | Monitoring | | | | Records | | | |

Figure 5

Within - teacher design: Each of the 4-level factors is described by two 2-level factors. The highest order interaction is used to divide the 64 cells into two balanced halves, - and +. The - half is then divided into eight blocks of four cells each. Each block describes the training sequence for one school.

| BLOCK | TIME | AREA | METHOD |
|---|---|---|---|
| 1 | Oct | Feedback | Traditional |
| | Jan | Evidence | Resource |
| | Feb | Records | Demonstration |
| | Apr | Monitoring | TV |
| 2 | Oct | Records | Traditional |
| | Jan | Monitoring | Resource |
| | Feb | Feedback | Demonstration |
| | Apr | Evidence | TV |
| 3 | Oct | Monitoring | Demonstration |
| | Jan | Records | TV |
| | Feb | Evidence | Traditional |
| | Apr | Feedback | Resource |
| 4 | Oct | Evidence | Demonstration |
| | Jan | Feedback | TV |
| | Feb | Monitoring | Traditional |
| | Apr | Records | Resource |
| 5 | Oct | Records | Resource |
| | Jan | Monitoring | Traditional |
| | Feb | Feedback | TV |
| | Apr | Evidence | Demonstration |
| 6 | Oct | Feedback | Resource |
| | Jan | Evidence | Traditional |
| | Feb | Records | TV |
| | Apr | Monitoring | Demonstration |
| 7 | Oct | Evidence | TV |
| | Jan | Feedback | Demonstration |
| | Feb | Monitoring | Resource |
| | Apr | Records | Traditional |
| 8 | Oct | Monitoring | TV |
| | Jan | Records | Demonstration |
| | Feb | Evidence | Resource |
| | Apr | Feedback | Traditional |

Figure 6

Rearrangement of within--teacher plan to show sequence in which area/ method combinations occur in each block. Each school in the between-teacher design is assigned one of the blocks.

presentation such as traditional in-service training, television equipment, etc?

G, H by J, K
(df = 6)

. . . in different areas, when different training methods are used; is there any evidence that some areas require certain training methods?

L, M Time
(df = 3)

. . . at different times in the school year?

G, H by F
(df = 3)

. . . in different areas, for teachers whose students performed about average versus those whose students did better than expected?

G, H by E
(df = 3)

. . . in different areas, for teachers with more structured versus less structured programs?

J, K by F
(df = 3)

. . . with different methods, for teachers whose students performed as expected or those

| | |
|---|---|
| | whose students did better than expected? |
| J, K by E (df = 3) | . . . with different methods, for teachers with more structured versus less structured programs? |

These hypotheses require 27 of the 48 degrees of freedom available, leaving 21 degrees of freedom for an estimate of residual error variance. As can be seen, numerous interactions of interest can be tested with this design. Most of the hypotheses can be profitably broken down into more precise questions in which specific areas and methods are compared. This would be a more powerful analysis than the omnibus questions presented in the table.

The methodology used in these experiments differs in several ways from that used in most traditional educational research. The experimentally controlled variations proposed here are designed to compare the effectiveness of several plausible alternative methods of training and different training practices. Most traditional experiments have compared an "experimental" treatment to a no-treatment control or a "business-as-usual" control. In the studies proposed here, variations in the targeted training areas and in the methods of training are designed to isolate the effects of specific teaching skills and of training methods used to promote acquisition of these skills.

The degree of control achieved by these designs is impressive. Differences between teachers can be handled by the design virtually to

the practical limit of our knowledge of factors affecting teacher effectiveness. The major hypotheses about treatment effects are all within-teacher questions. Each teacher serves as his own control, which allows highly sensitive tests of the treatment variations. Finally, notice the implicit assumption that these areas of classroom practice are largely independent of each other. We are assuming that a teacher's effectiveness can be increased by learning skill A, regardless of whether some other skill B has been adopted or not. Undoubtedly there are A's and B's which are not independent, but as a starting assumption for designing experiments, independence has the advantage of simplicity. According to this assumption, the effectiveness of a teacher in a classroom is a simple additive combination of the number of skill areas in which the teacher is proficient.

## In Closing

These proposals offer an alternative to present practices in curriculum research and evaluation, an alternative that is workable and may have considerable promise. It would reduce, if not eliminate, the distinction between formative and summative evaluation, a distinction which has often justified sloppy research during the formative stages of curriculum development, and largely irrelevant evaluation of the final product.

The cost of applying experimental design techniques to curriculum research is probably not much more than is being spent on curriculum evaluation in many federally sponsored labs and centers. What is required is a more active and analytic job of thinking by researchers during th rly stages of curriculum development. Rather than waiting

until the curriculum is finished and then wheeling out a battery of standardized tests, researchers would have to roll up their sleeves and work with developers during creation of a curriculum. The production of variant forms of a curriculum program to fit a design structure, the installation of these forms in carefully selected classrooms, the continuous assessment of these pilot versions through observation and testing, all entail the replacement of current trial and error procedures with a more systematic approach. The cost in dollars of using experimental designs would be relatively modest; the cost measured in careful thinking, precise impositions and systematic measurement would be considerable.

The benefits seem obvious. At worst, we would obtain trustworthy evidence to support the claims of skeptics that it really doesn't matter very much what goes on in the schools. The more optimistic hope is that by shedding light on the complex set of factors that make up an instructional program, we would see more clearly the differences between those practices that promote learning and those that do not.

## Footnotes

1. Paper presented to the Curriculum Symposium at the University of Delaware. This research was sponsored in part by a grant from the Carnegie Corporation of New York. I am grateful to Adrian Sanford and Annalee Elman for their comments, and to Jana Floyd, Frederick McDonald, Patricia Elias, and Kathryn Hoover for helpful discussion on this topic.

2. This example springs from a collaborative project with Frederick McDonald at Educational Testing Service.

## References

Anderson, N. H. Functional measurement and psychophysical judgment. Psychological Review, 1970, 77, 153-170.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook of formative and summative evaluation. New York: McGraw-Hill, 1971.

Bloomfield, L., & Barnhart, C. L. Let's read: A linguistic approach. Detroit: Wayne State University Press, 1961.

Bond, G. L. & Dykstra, R. The Cooperative research program in first-grade reading instruction. Reading Research Quarterly, 1967, 2(Whole No. 4).

Bruner, J. S. Toward a theory of instruction. Cambrige, Mass.: Belknap Press, 1966.

Calfee, R. C. Information-processing models and curriculum design. In Educational Technology, 1970, 10, 30-38. Reprinted in R. W. Burns & G. D. Broods (Eds.), Curriculum design in a changing society. Englewood cliffs, New Jersey: Educational Technology Publications, 1970.

Calfee, R. C. Sources of dependency in cognitive processes. Cognition and Instruction: 10th Annual Carnegie-Mellon Symposium on Cognition 1974, in press.

Calfee, R. C., & Floyd, J. The independence of cognitive processes: Implications for curriculum research. In Cognitive Processes and Science Instruction. Bern: Hans Huber, 1972.

Carroll, J. B. Psychometric tests as cognitive tasks: A new "structure of intellect." Paper delivered at LRDC Conference on the Nature of Intelligence, University of Pittsburgh, 1974.

Chase, W. G. Visual information processing. New York: Academic Press, 1973.

Cohen, E. G. A new approach to applied research: Race and education. Columbus, Ohio: Charles Merrill, 1970.

Cronbach, L. J. Evaluation for course improvement. Teachers College Record, 1963, 64. Also in R. W. Heath (Ed.) New Curricula, New York: Harper & Row, 1964, pp. 231-248.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, J. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J., & Snow, R. E. Individual differences in learning ability as a function of instructional variables. Final Report, Office of Education, Contract OEC 4-6-061269-1217.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-401

Haber, R. N., & Hershenson, M. The psychology of visual perception. New York: Holt, Rinehart, and Winston, 1973.

Kavanaugh, J. F., & Mattingly, I. Language by ear and by eye. Cambridge, Mass.: M.I.T. Press, 1972.

Kirk, R. E. Experimental design procedures for the behavioral sciences, Belmont, Calif.: Brooks/Cole, 1968.

Kirst, M. W., & Walker, D. F. An analysis of curriculum policy-making. Review of Educational Research, 1971, 41, 479-509.

Lindsay, P., & Norman, D. Human information processing. New York: Academic Press, 1972.

Morrison, D. F. Multivariate statistical methods, New York: McGraw-Hill, 1967.

National Bureau of Standards. Fractional factorial experimental designs for factors at two levels. Applied Mathematics Series 48. Washington, D. C.: Government Printing Office, 1957.

Rao, C. R. Linear statistical inference and its applications. New York: Wiley & Sons, 1965.

Scheffe, H. The analysis of variance. New York: Wiley & Sons, 1959.

Scriven, M. The methodology of evaluation. In R. E. Stake (Ed.), AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

Squire, J. R. How publishers develop effective and usable instructional materials in reading. Paper delivered to International Reading Association, New Orleans, 1974.

Stake, R. E. AERA mono, aph series on curriculum evaluation. Skokie, Ill.: Rand McNally & Company, 1967.

Sternberg, S. The discovery of processing stages: Extensions of Donders' method. In W. G. Koster (Ed.), Attention and performance II. Amsterdam: North Holland Publishing, 1969.

Stufflebeam, D. L., The use of experimental design in educational evaluation. Journal of Educational Measurement, 1971, 8, 267-274.

Suppes, P. The place of theory in educational research. In Educational Researcher, 1974, 3(6), 3-10.

Taba, H. Curriculum development: Theory and practice. New York: Harcourt Brace & World, 1962.

Tyler, R. Basic principles of curriculum and instruction. Chicago: University of Chicago Press, 1949.

Walker, D. F. What curriculum research? Journal of Curriculum Studies, 1973, 5, 58-72.

Walker, D. F. & Schaffarzick, J. Comparing curricula. Review of Educational Research, 1974, 74, 83-111.

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

Wittrock, M. C., & Wiley, D. E. The evaluation of instruction: issues and problems. New York: Holt, Rinehart & Winston, 1970.