CSE Report No. 28

THE DESIGN AND ANALYSIS OF EVALUATION STUDIES

David E. Wiley

Center FOR THE Study of Evaluation OF INSTRUCTIONAL PROGRAMS

University of California, Los Angeles, May 1969

THE DESIGN AND ANALYSIS OF EVALUATION STUDIES

David E. Wiley

University of Chicago

From the Proceedings of the

SYMPOSIUM ON PROBLEMS IN THE EVALUATION OF INSTRUCTION

University of California, Los Angeles
December, 1967

M. C. Wittrock, Chairman

Sponsored by the Center for the
Study of Evaluation

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

CSE Report No. 28, May 1969
University of California, Los Angeles

THE DESIGN AND ANALYSIS OF EVALUATION STUDIES

David Wiley

The original intent which motivated this paper was a desire to say something systematic concerning the design and analysis of evaluation studies. It seemed, however, that before anything explicit could be said on this topic, some clarity would have to come to my mind concerning the definition of evaluation and its important components. As a consequence, and at the risk of infringing on some of my colleagues' territory, I have set down some of my thoughts on this topic also.

The first section of the paper relates some arbitrary distinctions which I impose among the terms evaluation, assessment and appraisal. In the second, with the help of some previous work, I further narrow the definition of evaluation implied in the first section. The next section outlines and defines four separate components or elements of evaluation which seem to have been confused in some "evaluation" studies. The fourth and fifth sections concern three of these elements and their relations to certain notions about the design, analysis and measurement aspects of evaluation.

I. Evaluation, Assessment and Appraisal

The terms evaluation, assessment, and appraisal are often used interchangeably in attempts at gathering information about schools

or pupils. This is unfortunate, since the goals of information gathering and the methods used differ so widely. It seems worthwhile to attempt to give some guidelines for the uses of these terms and to explicate some of the similarities and differences implied by these guidelines.

Certainly, it would seem clear that there is some element of judgment either central to these processes or lurking in the background. That is, there is some "valuing" going on somewhere. If we look to one of our strongholds of meaning, the Webster's New Collegiate Dictionary (1953), we find a distinction between evaluation as "ascertainment of value," and assessment or appraisal as "setting of value." From these definitions I would place the judgmental or valuing element further in the background for evaluation, more related to assessment and appraisal. It might be reasonable to place the focus of "evaluation" on the process of ascertaining the levels of particular traits which are viewed as valuable, rather than on establishing which traits are valuable. Certainly the values must be decided upon and structured in some reasonable way before they may be ascertained for a particular object, but we may reasonably separate the two processes.

In this context, then, I would use assessment or appraisal for the general processes of judging what is valuable and ascertaining the particular levels of valued traits, while reserving evaluation for the latter.

We can now turn to the problem of what may be evaluated, assessed, or appraised. Harris (1947) makes some useful distinctions in his paper on school appraisal. He designates three aspects of the school which may be appraised: plans, resources, and processes. For Harris, plans consist of the goals of the school; resources are both physical and human, and processes are the activities--both instructional and noninstructional--which go on in the school. In a later paper (1963), he insists that the appraisal become evaluation only if the criteria consist of measures of pupil behavior. That is, in the context of the above discussion, the ascertainment of levels of valuable traits must be empirical and behavioral.

II. A Definition of Evaluation

In order to discuss the design and analysis of evaluation studies we must first define evaluation more explicitly. Two papers have been most influential in structuring my thoughts about evaluation. They are those of Harris (1963) and Cronbach (1963). Harris defines evaluation as

> ....the systematic attempt to gather evidence regarding changes in student behavior that accompany planned educational experiences.

Cronbach, however, places some differing emphases:

> ....collection and use of information to make decisions about an educational program.

We can see that the differences between the two definitions reside in the fact that Cronbach emphasizes the decision aspect of evaluation

while Harris emphasizes the behavioral nature of the criterion. Both emphasize the gathering of (empirical) evidence or information about planned educational programs.

In order to focus the content of this paper, I have taken the liberty of combining the two definitions into one which further narrows the concept of evaluation:

> Evaluation consists of the collection and use of information concerning changes in pupil behavior to make decisions about an educational program.

Thus I will be concentrating on behavioral information which is relevant to decisions about educational programs.

In addition to narrowing the evaluation concept to the information gathering or ascertainment process, a further comment is in order about the term "educational programs." In the discussion which follows I will restrict the meaning of "educational programs" to "instructional programs" or "procedures" and their components. This will tend to simplify the terminology in the rest of the paper, even though much of what I say will be relevant to other aspects of the school, e.g., the plans and resources of Harris (1947, 1963).

To summarize I will be concentrating on instructional programs or procedures and their components evaluated by means of pupil behavior.

III. The Elements and Terminology of Evaluation

In order to help conceptualize the evaluation process it is useful to distinguish certain elements by labeling them with special

terms. To these elements I have given the names: Standards, Objects, Vehicles, and Instruments.

The Standards of evaluation are a function of valuing or judgment process discussed earlier. The standards consist of the designation of traits which are considered important to evaluate (are valued), and the designation of levels of these traits which are considered desirable. A rough example of standards might consist of the statement that 90 percent of the pupils in a school should be able to read 70 percent of the material in a daily newspaper with 95 percent comprehension. This is incomplete since some of the terms in the sentence are not defined.

The Objects of evaluation are the instructional programs or procedures and their components. These might consist of something as complex as a "new math" textbook series or at the other extreme a particular "frame" in an auto-instructional sequence.

The Vehicles of evaluation are the carriers of the effects of the objects. That is, the pupils, classes, or schools.

The Instruments of evaluation are exhibitors of the behaviors of the vehicles. The selection or construction of these instruments is highly dependent on traits established as important by the standards, the particular objects to be evaluated, and the vehicles which are affected by those objects. In addition, the instruments used in an evaluation may interact with the standards, since the trait levels

considered desirable may be differently reflected in different instruments. Examples of instruments run the usual broad gamut of stimuli used for eliciting behavioral responses.

The main problem of evaluation, then, is to establish the effects of the objects on the vehicles by means of the instruments. The other element of the process is to compare these effects with the standards. The latter comparison will not be discussed in this paper.

IV. The Objects of Evaluation and Their Description

In order to evaluate an object (educational program, procedure, or component thereof) we must describe it, or at least be able to distinguish it from other possible objects. In its simplest form this description indicates the presence or absence of the object. Thus in the evaluation of complex educational programs such as textbook series we usually characterize the object by a dichotomous variable indicating its presence (or absence). Studies which involve the characterization of a complex educational program by its presence or absence are called "summative evaluation studies" by Scriven (1965). Two basic types of studies have been proposed by Cronbach (1963) to accomplish the goal of summative evaluation. These types he terms the "horse race" and the "time trial."

The "horse race" is the educational comparative experiment. In this procedure several different objects (educational programs)

are compared by randomly assigning relevant vehicles of evaluation to them and then comparing their standing on measures produced, by applying relevant instruments to the vehicles. The analytic procedure used for data generated in this fashion is usually the analysis of variance. This procedure transforms the dichotomous variables that indicate presence or absence of the various treatments into contrast variables that differentiate the treatments; then it relates them to the outcome measures.

The "time trial" is a procedure which ascertains the levels of the outcome measures in the presence of the object. It is mainly useful when one is not directly interested in comparison with other objects or when the conditions of the study can be assumed to remain constant. The trouble with the "time trial" study is that one is almost always interested in a comparison with some other objects, for if one were not, a decision would not need to be made. And given that comparison is necessary, the constancy of conditions becomes extremely important and is difficult to guarantee without the important concomitants of a comparative experiment. It is important to acknowledge, however, the pertinence of Cronbach's (1963) point that it is difficult to implement valid comparative experiments.

Another type of evaluation procedure defined by Scriven involves a different class of objects of evaluation. These objects are component parts or procedures of complex educational programs. The

general purpose in evaluating them is to gain information which will aid decisions about the modification or deletion of these parts from the overall program. Scriven terms this type of evaluation "formative." It seems generally accepted that the kind of study used in this type of evaluation should be both comparative and experimental. The analytic models used for the analysis of data from these studies are essentially the same as the analysis of variance models described above. An example of this kind of study might be an experiment comparing the effects of varying the sequence of certain instructional units or blocks within a complex instructional treatment. Another might be an optimization study of a particular unit of instruction.

There is a third type of evaluation study which seems to be little discussed. This type might be described by the phrase "making summative evaluation studies formative." This type of study involves making the description of the objects a quantitative characterization of the relevant traits of those objects and then relating that description to the outcomes.

An example of such a quantitative description might be the percentage of time a particular educational program spends with supplementary material as opposed to the basic textbook. This variable might be related to transfer objectives of the instruction.

This procedure may also be used to establish the effects on the vehicles of those characteristics of the instruction which are left

free to vary by the program. Note that in this case the emphasis has changed from the educational program (object of evaluation) to variations in the program (objects of evaluation). An example of this type of study was recently conducted by one of Benjamin Bloom's students at the University of Chicago (Anthony, 1967). An appropriate analytic model for this type of study is the standard regression model or appropriate modifications thereof.

V. Vehicles and Instruments of Evaluation

The appropriate vehicle for evaluation is highly dependent upon a characterization of the object of evaluation. The selection of an appropriate vehicle is equivalent to the selection of a sampling unit for a study.

To make this more concrete, if the object of evaluation is a typical classroom instructional program where the instruction is received simultaneously by all students in the class, then the appropriate vehicle (or sampling unit) is the class and not the individual pupil. This is equivalent to the standard definition of the experimental unit in this case: if two pupils in the same class may not receive different instructional treatments the classroom is the appropriate unit (see Page, 1965). Another way of looking at this is to say that traditional instruction is by nature classroom-based since if it were not it would be tutorial. This concept has been discussed by Wiley (1967).

The other vehicles are, of course, appropriate for objects of evaluation with other characteristics. Thus if one is conducting a formative evaluation study within the context of a computer-assisted instructional system, individual pupils may receive different treatments. The method is essentially tutorial, so the relevant vehicle (sampling unit) is the individual pupil. And if one is studying the effects of administrative policy the vehicle would be the collectivity supervised by the individual policy maker.

When the appropriate vehicle for evaluation is a collectivity, such as a classroom or school, a number of options open up with respect to instrumentation, the production of measures, and strategies of data analysis.

In the first place the measures produced by the instruments characterize the collectivity and not the individual pupil. This implies that not every pupil need be measured. That is, we may sample pupils from the unit and still be able to measure the status of the relevant unit. We may thus use completely random, or stratified random sampling schemes; and our only concern need be the reliability of the resulting measure with respect to the relevant unit.

It may be useful at this point to sketch a model for the reliability analysis of collectivity data in a simple case. If we are trying to differentiate reliably among classrooms by means of the

average scores on a test given to the individuals in that class, an appropriate model might be the following:

$$Y_{ij}=\mu + \alpha_i+\varepsilon_{ij} \text{ , where } i = 1,\ldots, n \quad j = 1,\ldots, m_i$$

n being the number classes and m_i the number of people in each class for which scores are available. The term ε_{ij} represents variation among individuals within classes and may be treated as error of measurement with respect to the determination of $\mu+\alpha_1$, the population mean score for the i^{th} class.

The reliability of the measure with respect to differentiating among classes is the proportion of variance in the mean score for each class attributable to true variations among classes. If we let $\sigma_\alpha^2 = \text{Var}(\alpha_i)$, and $\sigma_\varepsilon^2 = \text{Var}(\varepsilon_{ij})$ then the reliability of the i^{th} class mean is

$$\rho_i = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{m_i}}$$

This is equivalent to the intra-class correlation stepped up with the Spearman-Brown formula for an increase in test "length" of a factor of m_i.

An appropriate estimate of ρ, when m_i are equal is

$$\hat{\rho} = \frac{MS_\alpha - MS_\varepsilon}{MS_\alpha}$$

Bock and Wiley (1968) have determined that in a relatively homogeneous set of suburban school districts approximately 70 percent of the variation in the scores of a group of common standardized achievement tests is due to within-class variation. This would imply that the reliability of these kinds of tests with respect to differentiating classes would be approximately

$$\rho = \frac{.3}{.3 + \frac{.7}{30}} = \frac{.3}{.323} = .93$$

if the classes contain thirty pupils.

In other contexts, it has been my experience that measures of many traits, with respect to classroom means as observations, tend to correlate above .90 with measures of similar traits, implying that the actual reliability of standard achievement instruments for individual differences among classes is somewhat above .90. This would seem to be consistent with the Wiley and Bock data. For many purposes then, it would seem that samples of pupils would be adequate.

Another consequence of the unit being a collectivity is that each pupil does not have to receive the same items if the instrument

is a test with more than one item. Procedures for giving different pupils different items are called item sampling procedures and are due to Frederic Lord (1962). They are mentioned in the evaluation context by Cronbach (1963). These procedures may be very useful in that a complex trait, possibly represented by a population of distinct items, may be adequately assessed for units by giving each pupil a small and distinct sample of items.

One such design which seems to have great promise in evaluation studies may be described in the following way. Suppose one has a test consisting of m items. Randomly select m pupils from each of n classes. Randomly assign one of the m items to each pupil under the restriction that every pupil in a class is to receive a different item.

Class	1				2					n			
Pupil	1	2	....	m	m+1	m+2	...	2m	....	(n-1)m+1	(n-1)m+2	...	nm
Item	1	2	....	m	1	2	...	m		1	2	...	m

Note that pupils are nested within class and that items cross classes but not pupils. We may formulate the following model for the mn observations.

$$Y_{ij(k)} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_i(k),\ i = 1,\ldots,n;\ j = 1,\ldots,m;$$

$k = 1,\ldots,nm$

(where k is completely determined by i and j)

The expectations of the mean squares are

$$E(MS\alpha) = n\sigma^2_{\alpha} + \sigma^2_{\alpha\beta} + \sigma^2_{\varepsilon}$$

$$E(MS\beta) = n\sigma^2_{\beta} + \sigma^2_{\alpha\beta}$$

$$E(MS\alpha\beta) = \sigma^2_{\alpha\beta} + \frac{n}{n-1}\sigma^2_{\varepsilon}$$

and the reliability of the mean score for each class is

$$\rho = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \frac{\sigma^2_{\alpha\beta} + \sigma^2_\varepsilon}{m}}$$

which may be approximately estimated by

$$\hat{\rho} = \frac{MS\alpha - MS\alpha\beta}{MS\alpha}$$

This design is due to Cronbach, et al. (personal communication, 1967).

Another advantage of a collectivity unit is that single items may be used to characterize the unit in a reliable fashion. And as Cronbach (1963) has mentioned, items are easier to interpret than total scores on tests made up of many items.

Single item scores have other advantages also. It seems apparent (Wiley & Bock, 1967) that a major portion of the variation among schools on measures of achievement may be explained by a single source of variation which may be interpreted as social clsss. It might be expected, then, that if one took the item scores for a small but highly homogeneous set of items (e.g., ten items testing the addition of two one-digit integers) most of the covariation among the items could be accounted for by a reliable measure of social class.

A reliable measure of social class could be obtained by using almost any measure of prior achievement or ability, since the collectivity is the unit we are considering. This is so since variation

in backgrounds of the pupils within the community associated with the school is error with respect to the determination of the school mean. Thus the mean ability of the pupils in the school is likely to reflect the social class of the community.

It would seem to be a good hypothesis that most of the remaining covariation after removal of the variation in social class would be due to variation in instruction. This might then imply that the principle component of the matrix of partial covariances (removing social class), would be a relevant criterion measure for the evaluation of instructional methods. One might expect that little covariation would remain after removing both social class and the principle component.

We have explored some of the consequences for instrumentation and measure generation when the relevant unit is a collectivity but there are others. It would seem that the objects of instruction might well affect other characteristics of a unit than the mean level of achievement. They might, in fact, affect the distribution of achievement in the collectivity. If this is true the moments of the achievement distribution might be used as criterion measures. It would seem that the first four moments would be directly interpretable. For example, if one Program or Object of evaluation tended to produce more homogeneous achievement as indicated by a small variance or logarithm of the variance, this would be directly interpretable.

Another relevant kind of measure which may be produced is a contrast among subpopulations. For example, if one computed the mean score for boys and girls in a classroom and used the difference as a criterion measure, the effects on the difference score may be interpreted as an interaction of the treatments (if any) and sex of pupil with respect to the original criterion measure. It is important, however, to realize that this measure should be looked upon as a characteristic of the class rather than inferring what the effects of the treatments would be if the sexes were segregated.

This treatment of subpopulations may be extended to more than one way of classification. Thus we might create four subpopulations: High Ability Boys, Low Ability Boys, High Ability Girls, and Low Ability Girls. The four mean scores would produce three contrasts in addition to the mean: (a) a sex contrast, (b) an ability contrast, and (c) a sex-ability interaction contrast. These contrasts may then be used as separate criterion measures with possibly insightful results.

One might note that in the above example the two ways of creating subpopulations differ in that one was the result of a discrete variable (sex) and the other was a result of a continuous variable (ability). When the contrast is the result of a continuous variable it may be considered to be a rough estimate of the regression coefficient of the original criterion variable with respect to the continuous variable.

This logic might lead one to the use of regression coefficients as new criterion measures for evaluating the differential effect of the treatments on individual pupils.

I hope that some of the ideas and suggestions presented above will be helpful to evaluators and evaluation researchers in clarifying the muddy field of instructional evaluation.

REFERENCES

Anthony, B. M. The identification and measurement of classroom environmental variables process related to academic achievement. Unpublished Ph.D. dissertation, University of Chicago, 1967.

Cronbach, L. J. Course improvement through evaluation. Teacher's College Record, 1963, 64, 672-683.

Harris, C. W. The appraisal of a school-problem for study. Journal of Educational Research, 1947, 41, 172-182.

Harris, C. W. Some issues in evaluation. The Speech Teacher, 1963, 12, 191-199.

Lord, F. M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 22, 259-268.

Page, E. B. Recapturing the richness within the classroom. (Paper read at AERA Convention, February, 1965.)

Scriven, M. The methodology of evaluation. February, 1965, pp. 1-58 (dittoed).

Webster's New Collegiate Dictionary. Springfield, Mass.: Merriam Company, 1953.

Wiley, D. E. Standard experimental designs and experimentation under school conditions. (Paper read at AERA Convention, Chicago, February, 1965).

Wiley, D. E. The design and analysis of experiments. Highlights of the preconvention institute: Research designs in reading. Newark, Delaware: International Reading Association, 1967, Pp. 9-16.

Wiley, D. E., & Bock, R. D. Quasi-experimentation in educational settings: Comment. The School Review, 1968, 75, 353-366.

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

(TOP)

ERIC ACCESSION NO.	
CLEARINGHOUSE ACCESSION NUMBER	
RESUME DATE	5-1-69
P.A.	
T.A.	
IS DOCUMENT COPYRIGHTED?	YES ☐ NO ☒
ERIC REPRODUCTION RELEASE?	YES ☒ NO ☐

001

100 TITLE
The Design and Analysis of Evaluation Studies: Comments and Suggestions.

101
102
103

200 PERSONAL AUTHOR(S)
David E. Wiley.

300 INSTITUTION (SOURCE) UCLA - CSE — SOURCE CODE

310 REPORT/SERIES NO. CSE Report No. 28

320 OTHER SOURCE Symposium on Problems in the Evaluation of Instruction, Dec. 1967 — SOURCE CODE

330 OTHER REPORT NO.

340 OTHER SOURCE — SOURCE CODE

350 OTHER REPORT NO.

400 PUB'L. DATE 5-69- | CONTRACT/GRANT NUMBER OEC 4-6-061646-1909

500 PAGINATION, ETC.

501

600 RETRIEVAL TERMS

601
602
603
604
605
606

607 IDENTIFIERS

800 ABSTRACT

This paper is concerned with the design and analysis of evaluation studies and the need for definition of evaluation and its components. Wiley makes distinctions among evaluation, assessment and appraisal, and narrows the definition of evaluation. He outlines the separate elements of evaluation which have been confused in some studies, and relates these elements to design, analysis and measurement of evaluation.

801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822