COMMENTS ON PROFESSOR WILEY'S PAPER ENTITLED "DESIGN AND ANALYSIS OF EVALUATION STUDIES"

Chester Harris

University of Wisconsin

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. 29, May, 1969
University of California, Los Angeles

COMMENTS ON PROFESSOR WILEY'S PAPER ENTITLED "DESIGN AND ANALYSIS OF EVALUATION STUDIES"

Chester Harris

We have come into the third day of this conference, and enough things have been said in various contexts to make it possible for me to point out some things that bear in general on Mr. Wiley's paper, but still more generally on the whole set of papers.

I think that the most important contribution that can be made at this point in the conference is to identify and enumerate what I regard as three critical issues in the design and analysis of evaluation studies suggested in these papers and discussions. The area of design and analysis is actively changing and developing, and most of us would be hard pressed to predict the extent to which these issues will be resolved or reformulated in the near future. The measurement problem in evaluation studies involves a situation in which we have an instructional package that is to be used with some group of human subjects, and then evaluated in terms of how good it is. This demands that we adopt some scheme for specifying what we mean by "good."

There appear to be three types of "goodness" for those who take the behavior of students as the relevant evidence. One is goodness defined as a level of performance; a second is goodness defined as change of performance in a specified direction; and a

third is goodness defined as change of performance in a specified direction to a specified extent. Buried here are the questions of which behaviors are relevant and whether the observations that are made can become bases for inferences regarding learning as a result of the instructional package. This is an issue which Dr. Gagné posed for us earlier in the session. These three attitudes imply somewhat different measurement operations for any chosen type of performance. Let us leave this with the further acknowledgment that in any study many different types of performance may be regarded as important dependent variables, and that the amount of work required to make preparations for an evaluation study may be extensive.

The reality that there may be relevant dependent variables also suggests that appropriate designs for evaluation probably should be multivariate. This is the first issue which I wish to identify, the issue of univariate versus multivariate dependent variable studies. My strategy is not to resolve the issue but merely to enumerate the factors involved.

Possibly the simplest design for an evaluation study is that which employs only one instructional package and attempts to assess its goodness for two or more categories or types of students. Here we employ stratifying variables: age, sex, intelligence level, residential region, etc., to define our groups of students, and

then compare and contrast the various student performances. The intent of such a study is primarily descriptive (though tests of significance often are run): to define the goodness of the instructional package with respect to specified groups. This is a fixed-effects model, with the chosen levels of the stratifying variables being the only ones about which information is gained. Here there arises an issue which I will describe by extending the design so that more than one instructional package is used. I assume that we may retain one or more stratifying variables as well, and thus have a reasonably complicated design. I will not, however, complicate it by introducing repeated measurements. Such a design has as its intent a comparison among instructional packages for various groups and sub-groups. I repeat that in practice this is a fixed model; for we seem absolutely unable to define a population of instructional packages, and, even if we could, to be quite unwilling to select at random a set of instructional packages to study. Instead, we select the packages arbitrarily and deliberately; this is a fixed effect.

A design such as this has limitations that are inherent in all hypothesis testing. Among them is the familiar problem posed by the reasonable assertion that no sharp hypothesis can possibly be true. Testing such a hypothesis is merely an exercise in testmanship since the outcome depends heavily upon the manipular flexibility of the test.

It is perfectly reasonable to assert that no two instructional packages can possibly have identically the same effect; thus the testing of the hypothesis that two or more such packages have the same mean effect can be viewed as relatively unimportant. This represents my attitude toward the decision theoretic approach which has been mentioned over and over again at this conference.

Those who criticize hypothesis testing urge that we use estimation procedures instead. The question of what kind of estimation procedure is useful here is an important one. Some interest exists in developing an analogue of response surface methodology for evaluation studies. It is an analogue, since the elements of instruction packages that can be identified often exist in only a few discrete rather than continuously ordered forms. This creates some problems with the statistics, but in time these problems may be made manageable.

The response surface design attempts to vary inputs (elements of instruction) to the end of identifying an optimum or maximum output performance. This is quite a different approach to evaluation studies. The choice of this approach as opposed to the more conventional fixed model constitutes a second important issue.

Let me raise a third issue which is often associated with a Bayesian point of view in statistics. The fact that we tend to interpret every study as if it were being done for the first time

should make us uneasy, even though we still can not agree on how prior information should be incorporated into our analysis. Actually, there often are relevant prior findings that remain unused.

I am reminded of how we behave in directing dissertations. We always insist on a summary of previous findings in an early chapter, but we would be horrified if the student tried to integrate them numerically with his findings. The issue here is the extent to which, in any evaluation study, the design and analysis will ignore all the possible prior distributions.

A modification in practice--namely, learning to take into account the prior information--might be the one that would most improve the design and analysis of evaluation studies.

objectively definable within the system of transformations used, but in general the questions so derived by the particular set of transformations I have been talking about deal with what we ordinarily classify as explicitly stated facts.

More recently, however, I have begun analyzing the syntactic constraints existing between sentences. These analyses seem to be leading to an ability to deal with questions commonly judged to be testing "knowledge of higher level concepts and more complex processes." Indeed, I seem to be getting the intuitively satisfying result that the traditional essay question has a generic kinship to the mundane short answer completion question. The two types of questions simply represent transformations operating at different levels in the syntactic structure of the discourse.

Many of Anderson's questions appear to fall within the classes derived by these transformations. But some of them also appear to represent transformations of an order that differs from any I had yet thought about.

What I am arguing, then, is that we need a theory of test writing and that until we have such a theory, the practical use of evaluation for the formation of public policy does not seem to me to be possible.

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

TITLE

100 Comments on Professor Wiley's Paper Entitled "The Design and Analysis of
101 Evaluation Studies: Comments and Suggestions."
102
103

PERSONAL AUTHOR(S)

200 Chester Harris

INSTITUTION (SOURCE) | SOURCE CODE

300 UCLA - CSE

310 REPORT/SERIES NO. CSE Report No. 29

OTHER SOURCE | SOURCE CODE

320 Symposium on Problems in the Evaluation of Instruction, Dec. 1967

330 OTHER REPORT NO.

OTHER SOURCE | SOURCE CODE

340

350 OTHER REPORT NO.

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

PAGINATION, ETC.

500
501

RETRIEVAL TERMS

600
601
602
603
604
605
606

IDENTIFIERS

607

ABSTRACT

800 In his concern with indentifying the critical issues of evaluation studies,
801 Harris focuses upon the issue of univariate as opposed to multivariate
802 dependent variable studies, the arbitrary selection of instructional
803 packages, the tendency to interpret every study as if it were being conducted
804 for the first time, and the need to give adequate consideration to prior
805 information.
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822