CONFIDENCE TESTING: AN EXPERIMENTAL STUDY

A THESIS

PRESENTED TO

THE FACULTY OF GRADUATE STUDIES AND RESEARCH

UNIVERSITY OF MANITOBA

IN PARTIAL FULFILLMENT

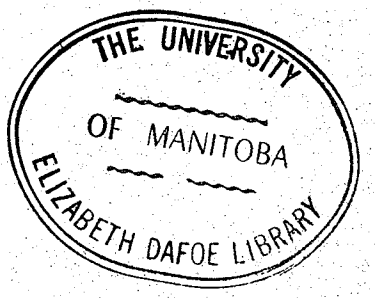
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF EDUCATION

BY

DORIS MOSS COWLEY

MAY 1970



ACKNOWLEDGEMENTS

The writer wishes to express sincere appreciation to the following for their various contributions:

Mr. W. Soprovich, Manitoba Department of Youth and Education, for facilitating data collection in the BSCS Blue Version pilot project schools; the students and their teachers at Churchill, Elmwood and St. John's High Schools and Fort Richmond Collegiate who took part in the study; Dr Peter A. Taylor for much helpful advice and constructive criticism.

ABSTRACT

The purpose of this study was to investigate a confidence testing procedure as a workable technique for the extraction of more information about a testee's state of knowledge than is possible under conventional testing procedures.

Confidence testing guarantees a testee that he may maximize his score if he weights his responses to each of the alternatives of a multiple-choice question in such manner as to honestly reflect his state of knowledge as to the correctness of each alternative. Confidence testing is claimed to have greater diagnostic utility than conventional procedures and by eliminating the need for guessing provides greater opportunity for the improvement of the teaching-learning situation and the psychological climate of testing.

Three hundred students comprised the sample. They were divided into an experimental, and four control groups, and were tested on 56 specially-constructed items on the BSCS Blue Version textbook in biology. Controls were imposed for test-taking instructions, scoring procedures, Blue Version biology content, and non-specific biology content. Analysis of the data obtained through student responses led to some insights into the confidence-testing method and to some tentative conclusions.

By comparing the experimental group's performance with the appropriate control, it was found that confidence-testing gave credit

for part knowledge; that testees found conventional testing procedures easier to follow than confidence testing procedures; that test items were biology-discriminative, though not necessarily Blue Version biology; and that the sex differential between boy-girl performance, which was clearcut under conventional scoring, was insignificant under confidence scoring. Girls exhibited a greater tendency to comply with confidence testing instructions.

Reliability of the test (.5) was low under confidence scoring, but was greater than that obtained when the same data were scored by conventional procedures (.4). Item-test reliabilities ranged from -.5 to .7 with relatively high standard errors of measurement. Test validity was also low (.4) and less than that obtained when the same data were conventionally scored (.7). The criterion selected was conventional school biology term-mark. These results were of the same order of magnitude as those found in other studies with confidence testing.

The items were found to be both difficult (82 per cent were greater than 50 per cent difficulty) and discriminating (only ten did not discriminate). Item characteristic curves were constructed for representative items.

It was concluded that confidence testing may serve as a useful diagnostic tool and that increased reliability and validity might be expected from specially-constructed items and a suitable criterion for

confidence-testing. Factors to be considered in the use of confidence testing and the interpretation of data are the homogeneity of the item and test content; homogeneity and ability level of the testees; item difficulty and discriminability; familiarity of the testees with confidence procedures and purposes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF DISPLAYS

CHAPTER 1

AN INTRODUCTION

1.1 Rationale.

> The aim of a good education in biology includes not only knowledge attained, both of the products and processes in biology, but the desire for knowledge and the ability to seek it. Hence, the energy of wanting, initially manifest in liking and respect for the teacher, must be shifted first from the teacher to the qualities the teacher possesses as an educated person. This energy must be shifted finally to the objects or materials of biological science. That is, the student must not only develop certain qualities and capacities in himself, but he must develop an interest in the subject matters of the major fields of knowledge that will cause him to continue to study them, pursue them, for the intrinsic pleasures of learning.

If one adopts a purposive view of behavior, the existence of a goal, or set of goals, is a necessary precursor to any teaching activity, though the precise statement of objectives in advance of work on materials and evaluation has not been deemed necessary by all developers of new curricula. The goals may be expressed as broadly as those above (Schwab, 1968, p.442), or the goals may be specified more behaviorally, that is, in terms of observable and measurable immediate behaviors. But if in the purposive view of behavior the existence of a set of goals is

necessary, mere existence is not a sufficient condition for determining a strategy for arriving at an end-product. Arrival at an outcome is validated only to the extent that process-data (formative evaluations) are positive evidence for the attainment of the goal.

The Biological Sciences Curriculum Study (BSCS) group, in attempting to impart the nature of biology as an investigatory science, has incorporated two broad aims into the course materials, each having implications for the kinds of learning expected of the students: substantive course content and scientific process. These two broad aims are interwoven with one another throughout all the course materials and together define the goal of student achievement in the BSCS context. Nine basic biological themes, all of which are represented with varying emphases in the three BSCS text versions, co-relate the content with the process aims.

As appropriate curriculum materials were developed, four objectives relevant to the BSCS philosophy emerged: three pertaining to the substantive content (memory, organization, and application of knowledge), and one pertaining to scientific process. In order for an achievement instrument to be valid in the BSCS context, therefore, each of these objectives must be taken into consideration.

The use of standardized objective tests has become an accepted evaluative practise in many schools and both standardized and classroom tests have an important bearing on the way students approach the

learning process, regardless of the subject-matter or the goals of the particular curriculum. Objective tests offer the decided advantage over essay-type tests that a teacher who is only partially trained in the skills of test-construction can look with some confidence to the reliability of the results from the test. Moreover, no matter how well-defined and desirable the objectives of a course of study may be, as far as the student is concerned, the key to success lies in mastery of the kinds of skills tested for. If tests require mere repetition of text detail then the student concentrates on rote-learning methods. Such a student must expect to go beyond mere recall in tests which demand the ability to apply knowledge to show relations and use skills.

One concern in the evaluation of a science is the disparity between philosophy and practice. Cohen (1957) observed that "our system of education tends to give children the impression that every question has a single, definite answer." BSCS materials encourage the student to discover that in many areas of scientific inquiry there is, in fact, no single, "right" answer but that some answers are either more, or less, correct than others because they differ in the degree of comprehensiveness. This being the case, items in which the alternatives vary in their degree of relevance yet are all plausible to the uninformed student, permit probabilistic responding that is appropriate to the philosophy of a non-deterministic science.

Regardless of the substantive context, one of the major purposes of testing at all is to provide formative ("feedback") data

upon which curriculum decisions can be effected. The normative use of standardized tests is inappropriate for this kind of decision-making since it ustilizes item responses averaged across people, ignoring the interaction of individuals with instructional strategies (materials, teachers, etc.). In order to assess this vital treatment effect, it is necessary to focus upon individual item-responses. A conventional testing strategy typically results in item scores which are either zero or one. Information is not obatined that could otherwise have been sought. A testee's part knowledge is disregarded. The testee is faced with conflict-situations as to whether or not to guess; he is encouraged to "outguess" the tester rather than be strictly honest about his state of knowledge; he is faced with a potentially large number of failure situations. A method which provides for the honest declaration of a state of knowledge thereby essentially eliminating troublesome problems of guessing and which, by its scoring system, motivates the respondent to give an honest response by allocating numerical credit for part-knowledge, contributes not only to the psychological climate of testing but also makes available a greater amount of item-information, thereby increasing the total utility of a testing program to an evaluator.

1.2 Purpose of the study.

The purpose of this study was to experiment with one particular testing strategy -- confidence testing -- which seemed to offer the

advantages of diagnostic utility to which conventional testing procedures do not lend themselves.

Since confidence-testing permits a testee to express his degree of belief in the correctness of a number of alternatives to a test item, the conventional multiple-choice item with its single "best" answer is not ideally suited to an experimental study of this kind. As a result, a set of test items was constructed which would provide the kind of response-setting from which the greatest number of inferences about the value of confidence-testing could be drawn. Because there was a need to make certain decisions about the merit of the BSCS Blue Version text and since the general framework of the BSCS materials seemed an appropriate medium upon which to carry out an experiment such as this, the items that were constructed were framed within the Blue Version context. The information yielded from responses to these items was available to those wishing to make assessments of the Blue Version text. Again -- the primary purpose of this study was to experiment with confidence-testing as an evaluative strategy and to make some judgments about its worth. A secondary payoff was that the information yielded by the specially-constructed test items could be used as feedback for any evaluative activities concerning the new biology program. No attempt was made subsequent to the study to employ the test data in the evaluative sense -- simply to delineate it.

CHAPTER 2

A SURVEY OF THE LITERATURE

The educational establishment is increasingly and continually confronted with a need to make decisions for which it has inadequate information. It is in order to meet this need that psychological and educational tests exist and that strategies of evaluation have received much attention over the past decade. Too often, testing has been equated with measuring the achievement of pupils in a normative framework, ignoring the need for information about instructional materials, teachers and administrators, the school and community environment, placement decisions, and interactions between each of these. The only real justification for the use of a test lies in its ability to provide information that will improve a decision-process beyond chance, or the base level. While any scientist realizes the utility of reliable information, the ultimate purpose of any measurement is to assist in the making of qualitative decisions that are in some sense "better" than those that would have been made on the basis of unaided judgement.

2.1 Decision theory.

One of the most significant developments in applied mathematics that has occurred since mid-century has been the conjoining of utility theory and probability theory to yield what is now generally referred to as decision theory.

Decision theory rose out of a concern for improving business and other economic decisions. Serious limitations were found in the theories of classical economics that emphasized the welfare of the individual entrepreneur. With the rise of megalithic business enterprises, there was an urgent need to consider how decisions are made by coalitions of subgroups with differing interests. The void in the theory of decision-making led von Neumann and Morgenstern to propose their Theory of Games (1947) in which a decision-maker was described as a participant in a game or a competitive market. This first attempt at describing decision processes proved to have value not only in economic, but also military, situations.

The publication of Statistical Decision Functions (Wald, 1950) extended hypothesis-testing into a general decision theory. The probabilistic framework of the statistician was applied to decisions in which risk comes from random variation of an event. Decision theory also takes into account the "utility" (benefit) of possible courses of action. The definition or estimation of such utilities constitutes much of the problematic nature of utility theory. Perhaps more than any other purpose to which statistical methods have been put, the determination of utilities has drawn together economists, psychologists, mathematicians and statisticians to account for an individual's choice of a course of action, that is, to determine what set of utilities is consistent with overt behavior. Decision theory, in principle, applies to all behavior and guarantees consistency between thought and action, without it being a moral system for dictating people's choices.