

CONFIDENCE TESTING: AN EXPERIMENTAL STUDY

A THESIS

PRESENTED TO

THE FACULTY OF GRADUATE STUDIES AND RESEARCH

UNIVERSITY OF MANITOBA

IN PARTIAL FULFILLMENT

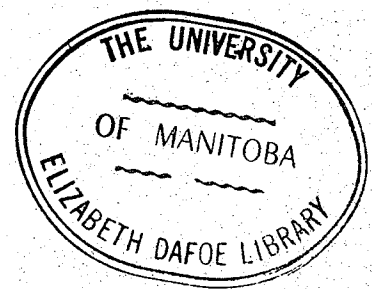
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF EDUCATION.

BY

DORIS MOSS COWLEY

MAY 1970



c Doris Moss Cowley 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

CHAPTER		PAGE
1	AN INTRODUCTION	1
1.1	Rationale	1
1.2	Purpose of the study.	4
2	A SURVEY OF THE LITERATURE	6
2.1	Decision theory	6
2.2	Utility	9
2.3	Subjective probability.	10
2.4	Degree of belief and exchangeability of events	15
2.5	Research on subjective probability.	18
2.6	Confidence testing.	19
2.7	The question of guessing	23
2.8	Summary statement	25
3	PROCEDURES	26
3.1	Generation of the item pool	26
3.2	Preliminary validation.	29
3.3	Assemblage of items.	30
3.4	Instructions for administration	30
3.5	Experimental design	32
3.6	Scoring procedures	33

CHAPTER		PAGE
3	PROCEDURES (Continued)	
3.7	Affective impact on testees	36
3.8	Analysis of results	37
3.8.1	Score-distribution parameters	37
3.8.2	Reliability (homogeneity)	37
3.8.3	Reliability (equivalence)	39
3.8.4	Test-criterion correlation ("validity")	40
3.8.5	Item-test intercorrelation	40
3.8.6	Item-criterion intercorrelation	41
3.8.7	Item difficulty	42
3.8.8	Item discriminability	42
3.8.9	Item characteristic curves	43
4	PRESENTATION AND INTERPRETATION OF RESULTS	44
4.1	Test parameters and their interpretation	44
4.1.1	Test parameters	44
4.1.2	Sex differences	49
4.2	Reliability	51
4.3	Validity	54
4.4	Affective response to confidence testing	54
4.5	Item analysis	57
4.5.1	Item difficulty	57
4.5.2	Item discriminability	60

CHAPTER		PAGE
4	PRESENTATION OF RESULTS (Continued)	
4.5.3	Item characteristic curves	60
4.5.4	Item-test intercorrelations.	60
5	SUMMARY, CONCLUSIONS AND RECOMMENDATIONS.	63
5.1	Summary	63
5.2	Conclusions	70
5.3	Recommendations	72
APPENDIX	73
	Table of item-scores	74
	Confidence instructions.	75
	Conventional true-false instructions	76
	Test items.	77
	Questionnaire	88
BIBLIOGRAPHY	90

LIST OF TABLES

TABLE		PAGE
1	Test parameters and significances of differences . .	46
2	Test parameters: significances of differences between boys and girls (confidence instructions).	50
3	Item characteristics.	58
4	Item-test intercorrelations and standard errors of measurement.	61

LIST OF FIGURES

FIGURE		PAGE
1	Item-difficulty as a function of test instructions	48
2	Item characteristic curves	59

LIST OF DISPLAYS

DISPLAY		PAGE
1	Calculation of item scores	35
2	Test reliability and validity.	53

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 utilizes 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 obtained 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.

Mathematical decision theory is a highly generalized theory, encompassing a variety of principles for decision-making. One such principle is that which attempts to avoid any sudden, large loss: the purchaser of an insurance policy on some good is functioning on this principle. Another principle would be that in which certain choices are preferred that have a small probability of accruing a large return: the purchaser of a lottery ticket would be acting out this kind of strategy. The most useful institutional decision strategy is that which maximizes the average gain (or minimizes that average loss) over many similar decisions.

Regardless of the principle employed, in order to evaluate a decision-making procedure, one is required to ask three questions (Cronbach and Gleser, 1957): does this decision seem the best possible on the basis of the given information?; would some additional information improve the decision process?; what difference is there in the goodness of decisions arrived at by two different procedures? What outcome results from a given decision is not only a function of the given treatment, but also of the characteristics of the individual and of a variety of situational variables. Thus, as Cronbach and Gleser point out (1957), any given test has a range of utilities depending on the use to which it is put. A validity matrix may be considered as an instance of the delineation of payoff, that is, of empirically determined utility.

2.2 Utility.

The basic premise underlying all utility theories is that one can assign numerical quantities to decision-alternatives in a way that determines that a particular alternative is chosen from a set of all possible alternatives if, and only if, the numerical quantity (utility) assigned is greater than that assigned to any other alternative. A general behavioral maxim is that an individual functions in such fashion as to maximize utility (Luce, 1967).

In order to maximize an expected gain it is necessary to make an assumption about the utility scale on which the decision outcomes are evaluated. It must be assumed that the value of each possible outcome can be expressed in equal units of "satisfaction", and that the units are additive. One might express utility in terms of a dollar scale, for example. If we think of risk as involving an assignable set of probabilities that sum to unity, then the utility of a risky alternative is the sum of the utilities of its component outcomes, each weighted according to the probability of its occurrence.

In contrast to risk, we can identify the concept of uncertainty. Whereas the probabilities associated with the determination of risk follow the conventional mathematical rule that their sum shall be unity, the probabilities associated with a state of uncertainty do not have this boundary restriction. For example, the probability of winning a dollar on the outcome of the toss of a coin (the risk of not winning a dollar) is $p = 0.5$. The uncertainty of winning a dollar will in this case be equal

to the risk because there are only two possible outcomes. An instance where uncertainty -- not risk -- is involved will entail the imposition of a subjective estimate of probability. Thus, betting on the outcome of a horserace would, in most circumstances, exemplify the principle of uncertainty. A comparison between the bets placed on each horse may be interpreted as an individual's evaluation of the likelihood of certain events occurring. The numerical equivalents of these dollar quantities are referred to as "subjective probabilities".

2.3 Subjective probability.

The term "subjective probability" has had two referents in the course of its evolution. First, it was the name for a school of thought about the logical basis for mathematical probability (de Finetti, 1951, 1937; Good, 1950). Second, it was a name for a transformation on the scale of mathematical probabilities which is related to behavior (Edwards, 1954). If subjective probability is assumed to be different from objective (mathematical) probability, as for example in games of chance, then the term "subjective probability" is best used in the second, or psychological, sense. Other terms with the same meaning are "personal probability", "psychological probability" and "expectancy" (Thrall, 1954). From the psychologists' point of view, the study of subjective probability is of theoretical interest as well as practical importance, for it at once provides a novel method and a viable conceptual scheme for the investigation of learning, thinking, perception and decision-making.

The generic term "probability" has itself had at least three contextual meanings: empirical, logical and subjective. The empirical view is that probability statements make assertions about the real world. It identifies probability with the conceptual limit of a relative frequency: the probability that A is B is $p = \lim(f/N)$, where f is the observed frequency and N the total number of events. p is given contextual meaning. Braithwaite (1964) suggests a theoretical-model concept in which the probability that $A = B$ is a parameter of the model given empirical meaning by a "rule of rejection". The "rule of rejection" in a statistical context is the well-known hypothesis-testing method of Pearson, Fisher and Neyman. In this latter viewpoint, the establishment of the "truth" of a probability-statement depends on the results of a non-terminating empirical investigation, the outcome of which is said to be only "probable" (Neyman, 1952; Fisher, 1956).

An alternative view is the denial that probability statements are empirical statements at all. Keynes (1921), Carnap (1962) and Jeffrey (1956) each defend the view that probability represents a logical relation between a proposition and a body of knowledge. Their argument is that between one statement and another statement(or statements) representing evidence, there is one and only one degree of probability that the statement may have relative to the given evidence. This implies that a probability statement is logically true if it is true at all. Probability statements are therefore purely formal: given a statement S and a body of evidence E , there is one and only one real number p such

that it may be said that the probability of S relative to E is p .

The denial of this latter assertion is precisely what distinguishes the subjectivistic view from the logical view. In the subjectivistic view, the relation between a statement and a corresponding body of evidence is a quasi-logical relationship and the number-value attached to it represents a degree of belief. This numerical value is not uniquely determined. A given statement may have any probability value between zero and one assigned to it on the basis of the given evidence, according to the inclination of the person whose degree of belief that probability represents.

The subjectivistic theory of probability, however, is not an empirical psychological theory of degrees-of-belief. Confusion about this has arisen from experiments performed with a view to finding out whether people's degrees of belief are related in some hypothesized way to the corresponding theory (Cohen, 1957). These experiments tested people, not the theory: the object was to find out if people have rational behavior patterns according to the prescriptions of the theory, not to find out if the theory accurately describes the behavior of people.

Perhaps, then, one might best describe the subjectivistic theory of probability as a logical theory in the sense that only certain combinations of degrees-of-belief in related propositions are admissible. If a person has a degree of belief, p , in a statement S , then he "should" have a degree of belief $(1-p)$ in the denial of S . An attempt to justify

the proscriptive "should" can be made by appeal to the argument that degrees-of-belief can be measured by betting ratios. For example, to say that for a given individual the probability of the occurrence of S is .25 is to say that he is willing to bet three to one against its truth. If the individual assigns any lesser ratio he is bound to lose anyway. That is, if he "bets" \$.25 on S and \$.25 on the denial of S, the "book" is assured of a profit of \$.50, regardless of the outcome. To avoid having book "made against one" an individual should distribute his degrees of belief so that they obey the rules of the conventional calculus of probabilities. The possession of such a distribution is called coherence. The notion of coherence was first introduced by Ramsey (1926) who took overt behavior in choice situations as indicative of degrees of belief. For behavior to be coherent, no set of bets on a series of propositions was allowed which ensured that, no matter what the outcome, the nettor would lose. There is a logical demand that one be coherent in one's beliefs -- and this is the only demand made by the subjectivistic theory.

There are two current meanings attached to the term "coherence". The first meaning maintains the sense imposed by Ramsey and reaffirmed by de Finetti (1937) and Lehmann (1955). A second sense invokes the notion of "strict coherence" (Shimony, 1955; Kemeny, 1955) in which not only is it impossible for the holder of coherent beliefs to lose but also it is impossible for him to place bets so that he will not win some zero-amount (that is, come out even) and there is a chance

that he will suffer a net loss. Thus in the older, weaker definition, a bettor is said to be incoherent in his behavior if he distributes his degrees of belief in such a way that he must lose, while in the case of the strong definition, the bettor is said to be incoherent if he distributes his degrees of belief in such a way that he may either come out even or lose.

In the subjectivistic theory, probability represents the degree of belief that a given person has in a given statement on the basis of given evidence. A person should be consistent in the strict logical sense. In fact, some writers use the term "consistency" in place of "coherence" (Edwards, 1954). It is perhaps preferable in referring to matters of judgment that the term "consistency" keep its lay meaning, in the sense that beliefs would be deemed consistent to the degree that they do not contradict each other.

By the rules of deductive logic, if the evidence logically entails E, then the individual should have the highest degree of belief in E; if the evidence entails the denial of E, he should have the lowest belief in E. The subjectivistic theory of probability goes one step further: it posits that a person's body of beliefs, considered as a whole, must be coherent as well as consistent. Such theories are subjectivistic or personalistic in the sense that an individual may hold any degree of belief in any given statement on any given evidence, provided only that his degrees of belief in other, related, statements are suitably adjusted (Davidson, Suppes and Siegel, 1957; de Finetti,

1961; Frechet, 1954, 1955). In either of the alternative views (the empirical and the logical) there is one and only one degree of probability that can be assigned correctly to a statement relative to a given body of evidence. The subjectivistic view encompasses any degree of belief in any statement but restricts the distribution of degrees of belief among related statement-sets.

2.4 Degree of belief and exchangeability of events.

If the physical universe is regarded as deterministic in structure, complete description of nature involves knowing all the true statements about the causal relations between events and thus being able to predict with certainty the future course of nature. In such a scheme a statement of probability reflects a level of ignorance. Bernouilli (1713) was probably the first to define probability as a degree of confidence in a proposition whose truth was indeterminable. The degree of belief, or degree of confidence in a proposition, identified as its probability of occurrence, is a function of the knowledge that a person has at his disposal and may therefore vary from one individual to another, or from time to time. The "art of guessing" consists of estimating as precisely as possible -- on the basis of available knowledge -- the "best" values of probabilities.

De Morgan (1847) explicitly defines probability in terms of degree of belief. De Morgan's argument is essentially a reductio ad absurdum. For example: the infallible feeling in our "knowledge" that $2 + 2 = 4$, we call "certainty". If we treat "knowledge" as a magnitude

then we can gainfully talk about degrees of knowledge. Lower grades of knowledge-amounts are called "degrees of belief". Probability refers to, and implies, belief; thus belief is but an alternative label for imperfect knowledge.

Keynes (1921) suggested that "probability" is an undefinable logical relationship between one set of propositions and another set. More importantly, in context, he introduced the idea that this relationship is associated with the rational degree-of-belief in a proposition. Keynes assumed that all degrees of belief are neither measurable nor comparable: he thus avoids the difficulty of assigning some numerical equivalents to a degree of belief. Ramsey (1926) and Borel (1924) each independently proposed that the only sound way of measuring a person's degree of belief is by identifying the latter with specific kinds of overt behavior. If a person is willing to gamble on the outcome of the occurrence of rain tomorrow by tossing a coin, then one can say that his degree of belief in the proposition that it will rain is numerically .5. Koopman (1940), following Keynes, retained an intuitive notion of "degree of belief" and argued that such probabilities are not necessarily completely ordered, while measured degrees of the same thing must be.

The introduction of the concept of "equivalence", "symmetry" or "exchangeability" (de Finetti, 1931), was an attempt to bridge the notions of subjective probability and the classical procedures of Bayesian statistical inference. Bayesian procedures are a body of methods for inferring outcomes, according to which one starts with an a priori

probability distribution which is then modified in the light of experience and experimental evidence.

In the case of exchangeable events, certain types of inference are independent of the original assignment of probabilities to the individual events of a sequence. de Finetti (1964) shows that, for example, in the case of a sequence of exchangeable events a person, whatever his initial position, must, if he is to be coherent in his beliefs, come eventually to assign a probability to the event in question which is close to the observed relative frequency. The principle of exchangeability stresses that a sequence or order in which events occur has nothing to do with determining its associated probabilities: it is the observed relative frequency which is all-important.

The subjectivist would argue that events in a sequence are not independent, or are at least not viewed as independent in establishing knowledge of degrees of belief. Subjectively, the occurrence of certain events in a sequence suggests evidence about the occurrence of future events which affects our degree of belief about them. Thus, the subjectivist would say that the (conditional) probability attributed by an individual to the k th toss of a coin showing heads, given knowledge of the previous $(k - 1)$ outcomes is dependent upon the proportion of heads showing in these $(k - 1)$ tosses, yet is independent of the particular order in which heads appeared.

2.5 Research on subjective probability.

By 1955, all current decision-making models asserted that a decision-maker behaves as though he compares payoffs, and chooses that course of action from among those available to him for which the sum of the probability-utility product is greatest (Edwards, 1961). The models differed to the extent which they permitted objective measurement of utility. Measurement procedures for assessing an individual's subjective probability were developed by Toda (1963), van Naerssen (1961), de Finetti (1962) and Roby (1965), all having the property that an individual would maximize his expected utility if, and only if, he expressed honestly his subjective probabilities. Shuford, Albert and Massengill (1966) extended this notion under the rubric of admissible probability testing. Thus the argument had reached a point where, for expected maximum utility models to satisfactorily represent a state of nature, subjective probability would have to replace objective (mathematical) probability.

Some progress towards clarity about the nature of subjective probability resulted from Savage's (1954) work. Savage based his analysis on two assumptions: the assumption that all courses of action can be rank ordered for a given individual, and the assumption that if a course of action A is at least as "good" as a course of action B in all possible future states of nature and is definitely better in one or more states, then B should never be preferred to A (the "sure-thing principle"). On the basis of these assumptions, Savage defines "subjective probability" as a number that represents the extent to which an individual thinks that

the occurrence of a given event is likely. This number has the same mathematical properties as objective probability (Cohen and Hansel, 1958).

Apart from the efforts by Savage, and particularly by Cohen and his co-workers, there is little or no systematic evidence about the nature of subjective probability.

2.6 Confidence testing.

The point has been made thus far that any decision-making process is dependent upon determining the utility of an outcome and that the utility is, in turn, a function of the product of the value of an event and of the probability of its occurrence. The current view is that the best measure of utility is arrived at by invoking the measure of subjective probability and value.

Since evaluation -- which assumes such importance in the functioning of schools -- is but one application of decision-making, it too depends upon questions of value and upon measures of subjective probability. Insofar as evaluation depends upon classes of evidence derived from the administration of tests of various kinds it is compelling that the tests themselves must be couched in such fashion as to permit the expression of subjective probabilities. Conventional choice testing (that is, testing which employs items that involve choice from a number of stated alternatives) has typically restricted the selection of response to that one alternative which is judged "best" (most appropriate) by the testee, according to some criterion.

In contrast, test response procedures that require an expression of confidence of the testee in one or more of the given alternatives (confidence testing), permit the declaration of precisely those subjective probabilities that optimize decision-making for the test user. Confidence testing yields more precise information about a person's state of knowledge than conventional choice-testing does. This information can be used to improve the effectiveness and efficiency of selecting, classifying and training individuals.

Although, in the history of testing, it has been suggested that the use of confidence testing could greatly increase the amount of information available from the test, experimental attempts to measure confidence have generally been failures. Many of the earlier attempts to measure confidence were scored in such a way that, if the student had slightly more confidence in one alternative than in any of the others, he could maximize his expected score by pretending that he had complete confidence in that alternative. Or, if he had no knowledge of the item content, any confidence-response was as good as any other. In order for confidence testing to yield valid and reliable (stable) results, it is necessary to have a scoring system which makes it possible -- and in the best interests of the student -- for the testee to state honestly his degree of confidence whatever his state of knowledge, and thereby maximize his score. Current testing techniques for assessing student knowledge not only fail in this respect, but they also fail to extract all of the potentially-available information from the test responses.

To extract this additional information, "admissible probability measurement" procedures have been proposed. Commonly used admissible procedures are, for example, direct estimation, category judgments, direct ratio-scaling, and indifference procedures. Admissible measurement procedures include the quadratic scoring system which was proposed independently by Roby (1965), de Finetti (1962), Toda (1963) and van Naerssen (1961) and the reproducing scoring system (RSS) of Shuford et al. (1966). In each of these techniques, a testee's score is determined by the probability assigned by him to the correct (keyed) response plus his distribution of points (probabilities) to the other, non-keyed responses. Such scoring systems are given the generic title of "symmetric scoring systems" (SSS) because the numerical values assigned to the incorrect responses can be interchanged without affecting the total item-score. A general recommendation seems to be (Riphey, 1966) that if test items contain a single, correct response the RSS (spherical or truncated logarithmic forms) is the better scoring system; if the items have more than one possibly-correct response, the generalized quadratic (Euclidean) form is preferable. When a group of testees is highly homogenized, confidence-testing procedures have little advantage over conventionally-scored tests. In part, this is an artefact of the structure of correlation coefficients in general. In those instances when a tester has no information whatsoever about a group, scores resulting from choice-testing yield about half the information resulting from confidence-testing; the same situation obtains when students' states

of knowledge are broadly distributed with relatively few well-informed or misinformed students. With a uniformly uninformed group (such as a class of students starting out on a new subject) the gain from confidence testing over conventional testing is trivial, but, of course, in such circumstances the amount of information gained by any testing program is negligible.

Perhaps the greatest justification for confidence-testing in comparison with conventional choice-testing lies in the amount of diagnostic information resulting from the use of the scoring system. In a conventionally-scored test, each item is typically scored zero or one, with the total test score being the sum of the item-scores. For a test scored according to any of the confidence procedures an item score may range anywhere from zero to one on a continuum, the actual score being determined by the number of points in total to be distributed, by the number of points allocated to the keyed correct response and by the distribution of points across the remaining, unkeyed, responses. Because each item yields scores on a continuum, there is necessarily more diagnostic information available. Shuford et al. (1966) suggested five categories of states-of-knowledge assignable from continuum item scores: well-informed (a high degree of confidence in the correct response); moderately informed (a fairly high degree of confidence in the correct response); uninformed (equal confidence in all the answers); partially informed (high confidence in the correct response but the same degree of confidence in one or more of the incorrect responses) and misinformed

(low degree of confidence in the correct response, high degree of confidence in one or more of the incorrect responses). Not only may a testee be categorized in terms of his response to a given item but also his total test score will yield certain additional information. A total test score has importance because it has typically been used for determining such things as course grades, placement and selection. Because a test scored under confidence-testing conditions takes into account partial knowledge a different ordering of students may result from that arrived at by conventional scoring methods. The conventionally-scored test does not discriminate between partially informed, uninformed and misinformed students. Therefore, more valid decisions are possible on the information gained from confidence-testing.

2.7 The question of guessing.

Two things determine a student's score on a test: knowledge, and strategy. Knowledge is defined as the degree of confidence in each of the given alternatives (Shuford et al., 1965). Conventional choice scoring systems encourage the strategy (for the individual wishing to maximize his expected score) of not skipping an item, and if the testee does not have maximal confidence in a single alternative then he should arbitrarily choose from among those alternatives in which he has equal confidence. That is, in a test situation which is scored (0, 1) per item, it is in the best interests of the testee to respond to every item and to guess intelligently on those items for which he is unsure of the correct response. In the case where some correction for guessing is

applied, the optimal strategy for the testee is to omit the items about which he is unsure of the correct response, thus avoiding the penalty of making an incorrect guess. In sharp contrast, admissible probability procedures do not require the testee to make up his mind whether to guess, or not. His best strategy is to be completely honest: he should neither skip an item, nor guess, but declare his lack of information by assigning equal confidence-values to each of the given alternatives. According to any of the admissible probability scoring formulas, the resulting score will be approximately 0.5: conversely, a score of about 0.5 represents a state of being uninformed. Thus, item scores should be interpreted as a means of categorizing the state of knowledge of an individual about the item content. This interpretation makes provision for specialized instruction particularly obvious (Shuford et al., 1966). Further, since one set of test scores may be used for many different purposes, the extent of guessing will differentially affect the quality of the decision made. The admissible probability scoring system "flags" all those items about which the respondents have less than complete knowledge. If the proportion and/or extent of lack of information is judged to be unsatisfactorily high, the results of the test may be interpreted in some different fashion, or disregarded altogether. If the proportion of guessed responses per testee varies considerably, the reliability of group results might be questioned (Massengill and Shuford, 1966). None of this information is made explicit under the conditions of a conventionally-scored instrument.

2.8 Summary statement.

Unquestionably, the research on subjective probability has not been sufficient to make even an approximately final statement about its tentative usage in an evaluative situation. As educators' thoughts turn to a concern for an increasingly individualized educational process there is inevitably a need for more sensitive instrumentation to reflect individual states-of-knowledge. Two strong points follow from current views about the educational process: one is a tolerance for "telling it like it is" and the other is a greater technical potential for individual instruction. Admissible probability testing encourages honest responses without penalty for lack of information or for misinformation. The technological advantages -- at least potentially -- that have accrued in the past decade permit an instructor to correct for incomplete information. A third point, derivable from the philosophy of science, that can be advanced in favor of confidence testing, is that all knowledge is not finite. Any given problem is likely to have more than one solution and the critical behavior is the ability to make the best decision to resolve the problem, given a set of circumstances. Evaluative procedures that reflect the probabilistic nature of problem-solving are more allied to the realities of decision-making.

Contributions have been made through decision theory by focussing on the determination of utility and subjective probability, to a method of scoring test responses which essentially eliminates concern for guessing and allows increased payoff from test data.

CHAPTER 3

PROCEDURES

The notions of subjective probability clearly have to be tested in some material context. In this instance it was decided to work within the framework of the BSCS Blue Version text, which was being piloted in Manitoba schools with a view to possible adoption. The data generated would therefore have utility not only to a study of confidence testing but also for making evaluative decisions about the biology curriculum.

To this dual end, a set of biology items was generated which covered approximately one third of the content of the Grade 11 Biology course. The remainder of this chapter is devoted to an explication of the methodology employed in the creation and administration of these items and in the analysis of responses to them.

3.1 Generation of the item pool.

The content objectives of each chapter of the Blue Version biology textbook may be inferred from the Teacher's Guide which accompanies the text. The values of the BSCS Committee in regard to these objectives are such that each chapter of the text may be taken as having equivalent value to all other chapters. The task of writing items was therefore essentially one of reflecting the content of each chapter in approximately equal proportions.

Items were written using the multiple-choice format with five alternatives per item. The first step in item-writing was to identify the content of each chapter. A number (about 200) of items was then written, isomorphic with the chapter content and also reflecting the general underlying themes proposed by the BSCS group. A number of items not specifically related to text-content was also written. These latter items were designed to reflect general understanding of science. Items were constructed to differ in difficulty, in reading level, in type of thinking involved (analytic, synthetic, critical, etc.), amount of information required of the testee to respond to a given item. Each item was therefore considered as requiring essential content-information on the part of the testee and was contributing to a wholistic framework. It must be stressed that in this study, the biology content was being used only as a convenient medium for confidence-testing, so that issues concerning the merit of the biology curriculum are irrelevant.

Under conventional multiple-choice testing conditions, it would be essential to ensure that there was, indeed, one, single "best" choice. In the context of confidence-testing, it is no longer appropriate to be concerned with only one "best" response, but to have each alternative at least partially contributing to a solution to the stem problem so that the respondent may express his degree of confidence in the correctness of each of the available choices. All items in the item-pool were constructed within the constraints of the confidence-testing model, that is, that each alternative was constructed to appear plausible.

Following the usual proscriptions of item-writing (see, for example, Helmstadter, 1964), every effort was made to

- i. word each item-stem and alternative lucidly;
- ii. provide only enough qualifications in each stem to delimit the response basis;
- iii. avoid overlapping or inclusive alternatives except insofar as it was judged necessary for testing for the understanding of terms;
- iv. make each alternative grammatically consistent with the stem;
- v. avoid direct quotes or stereotypic technical phraseology;
- vi. avoid the use of specific determiners ("always", "never");
- vii. avoid position-cues for the correct alternative;
- viii. make each alternative about the same length;
- ix. avoid deliberately misleading questions;
- x. place the alternatives in logical order where one exists;
- xi. make all unkeyed responses plausible to persons who lack the required information.

With respect to making plausible alternatives, some questions, because of their content emphasis, were particularly useful in testing for partial knowledge, so that alternatives were specifically designed with overlap. In addition to a proportion of the alternatives overlapping, it was made clear to testees that some items had more than one alternative keyed as being correct. By raising such minor ambiguities

the potential for testing for partial information was maximized. Testees were also thereby encouraged to shake the psychological set of looking for but one correct alternative. Where some items were used to test for an ability to distinguish the correct response from partially correct and incorrect alternatives, other items were used to distinguish one totally incorrect alternative from among four correct statements. These latter questions employed "not" and "except" in the stem and on occasion a double negative resulted in the stem, and the alternatives. Although the occasional use of double negatives undoubtedly increases the reading difficulty level and is therefore not regarded as good item writing practise under most circumstances, because of the restrictions of producing items for confidence-testing, it was sometimes necessary to employ this device.

3.2 Preliminary validation.

The preliminary item pool which was composed of 200 draft items was subject to careful face-, sampling-, and content-validation in relation to the biology textbook, and to logical analysis for appropriateness for confidence-testing. Face-validation relied on a subjective evaluation of what the test appeared to measure. Although far from being a stable basis upon which to evaluate a test, face validity is about all there is upon which to rely during the initial stages of item writing. Sampling validity is also -- to an extent -- subjective in that it requires matching the items with behaviorally-stated objectives and breakdowns of the trait- or content-area to be measured. In this

way the tester is assured that the content domain is reflected in the appropriate ratios and the desired behaviors are appropriately examined.

Applying these preliminary validity criteria, a set of 56 items was selected. By happenstance this number of items was judged to be suited to the time-limitations of a typical class period.

3.3 Assemblage of items.

The 56 selected items were arranged in order of textbook chapter presentation. No attempt was made to prejudge the items in terms of their difficulty as a basis for item-arrangement. All 56 of the items were retained, not only because this number permits response within reasonable time limits, but also because the inclusion of a large number of items would contribute to the reliability of the item agglomerate.

3.4 Instructions for administration.

Two sets of instructions were generated.

Since this study was an experiment concerning confidence testing, one set of instructions pertained directly to this end. As in any strong experiment, a comparison group ("control") is necessary. In this case, test-taking behavior was to be controlled for: hence a second set of directions.

The directions actually given to both the experimental and the control groups are provided in the Appendix. For the group responding

according to confidence-testing procedures, each testee was provided with an answer sheet of the familiar IBM type, with five blanks, labelled a through e, corresponding to the five alternatives per item. Testees were told that they had ten points per item which they could distribute in any manner they chose across the five alternatives in order to indicate numerically their confidence in which of the given alternatives was/were correct. They were told that if they did not recognize any particular alternative as correct, they should distribute the ten points among the alternatives they could not eliminate. In the event that they could not eliminate any of the alternatives they were to indicate their complete lack of information by distributing the ten points equally among the five alternatives.

The second set of instructions asked the testees to decide if the response keyed as correct on their answer sheet was indeed correct. If they agreed that it was, they were to check the provided column; if they disagreed, they were to mark the appropriate column on the answer sheet and provide a corrected response. They were aware that these items were to be scored conventionally (0,1).

In both cases, testees were told that their responses to the test would in no way influence their school standings in biology.

3.5 Experimental design.

Three hundred students constituted the experimental sample. Forty-three students acted as substantive control; the remainder was approximately half of the population piloting the Blue Version material, of which forty students acted as instructional control.

The experimental group (E) was exposed to a content examination on the Blue Version materials with instructions to respond to the items according to confidence-testing procedures. These tests were scored by the reproducible scoring system (q.v.). Controls were provided on four levels. The first control (Cc) required a response from the testees according to the confidence-testing approach but these responses were scored conventionally (0, 1). This control allowed comparison between the total scores, using both the conventional and reproducible scoring systems. The second control (Ctf) was tested under conventional true-false instructions on the Blue Version content, and the responses were scored (0,1). This control provided a comparison of scoring method based on different instructions. It should be noted that the logical complement of this control (true-false instructions - reproducible scoring) is not empirically realizable. The third control (Cy) was a group taking biology but not the specific Blue Version content. A fourth control (Cp) was made up of students who had taken no biology. Both of these latter groups were required to respond probabilistically and were scored according to the reproducible system. These groups provided data by means of which substantive discriminations could be established.

3.6 Scoring procedures.

Tests were scored by one of two procedures: conventional (0,1) scoring and reproducible scoring (RSS).

The conventional scoring procedure allocated an item score of 1 for an item-response that coincided with the keyed answer, 0 for any other response. No correction for guessing was applied. Total score was obtained on the 56 items by summing the individual item scores. Under the instructions whereby a testee was required to distribute the ten points allowed for each item, the probabilistic response was scored conventionally by allocating one point to those responses in which the maximum point-allocation corresponded with the keyed answer. This is a rational system, since it might well be presumed that if he were permitted only one response, that response would have been given by the testee to that alternative in which he had most confidence. If more than one alternative was keyed as correct, a full mark was given where each keyed alternative was demonstrated to have had equal appeal to the respondent. A weighted average was calculated for other point distributions. Once again, a total score was found by summing over the 56 items.

The reproducible scoring system (RSS) involves the computation of item scores by means of one of a number of formulas. The formula first chosen was that derived by de Finetti (1965):

$$S = r_h + 0.5(1 + \sum r_j^2)$$

where S = item score

r_h = number of points assigned by the testee

to the response keyed as correct
 $\sum r_j^2$ = sum of the squared points assigned by
 the testee to each alternative.

Since there were ten points to be distributed per item, each point-distribution was multiplied by 0.1 to ensure that the maximum item-score was unity.

The formula finally chosen for arriving at item scores was the Shuford-Massengill (1966) spherical reproducing formula (S-RSS):

$$F = r_h / (\sum r_j^2)^{0.5}$$

where F is the item score and the other symbols retain the same meaning as those in the de Finetti formula above. The choice of the S-RSS was determined by certain anomalies that result from the de Finetti formula. One typical problem is that the score allocated to a person who assigns zero points to the keyed response and who assigns the ten points across more than one of the incorrect alternatives, is positive using the de Finetti formula, owing to the summative nature of the residuals (r_j 's). For the S-RSS, this problem does not arise, since $F = 0$ whenever r_h is set equal to zero. Examples of item scores from both formulations and the calculations pertaining thereto are given on the next page (Display 1). A complete table of all possible item scores according to all distributions of ten points, calculated on the basis of the S-RSS, is provided in the Appendix.

DISPLAY 1: CALCULATION OF ITEM-SCORES

Alternatives	Well- informed	Moderately informed	Uninformed	Misinformed	Misinformed
	r_j r_j^2	r_j r_j^2	r_j r_j^2	r_j r_j^2	r_j r_j^2
a	0 0	0 0	.20 .04	.50 .25	0 0
* b	1.0 1.0	.50 .25	.20 .04	0 0	0 0
c	0 0	0 0	.20 .04	0 0	1.0 1.0
d	0 0	0 0	.20 .04	.50 .25	0 0
e	0 0	.50 .25	.20 .04	0 0	0 0
Sum of r_j^2	1.0	.50	.20	.50	1.0
F	1.0	.71	.45	0	0
S	1.0	.75	.60	.25	0

Key:

* = keyed alternative

F = item-score calculated by R-SSS

S = item-score calculated by de Finetti formula

A number of options was open in the case where more than one alternative was keyed as being correct. For sure, a student was given full credit for recognizing the multiplicity of response and answering accordingly. Where his responses suggested that the testee did not recognize that several alternatives were desirable responses, the options were to provide bonus points or to average the score that resulted from the particular distribution of points to the keyed alternatives. To maintain a uniform total-score ceiling it was judged more reasonable, in context, to adopt the averaging procedures. For example, if two alternatives were keyed as correct for a given item the expected response would have been an allocation of five of the ten points to both of the keyed alternatives. This would have gained the full mark. If, however, the testee had placed all ten points on just one of the keyed alternatives, his score would have been 0.5, or the average of placing ten points and zero points on the keyed alternatives. Other distributions would have resulted in intermediate item-score values.

3.7 Affective impact on testees.

Teachers of each of the experimental groups were provided with copies of the students' scores which they were asked to communicate to their classes. Testees, in possession of their scores, were then asked to reply to a series of questions which were designed to assess their beliefs about the efficacy of the confidence-testing method. A copy of this questionnaire is provided in the Appendix.

3.8 Analysis of results.

The item responses from the experimental and control groups were subjected to the appropriate scoring procedure and these scores were then analyzed to yield a variety of information including score distribution parameters, indices of reliability and validity, difficulty and discriminability. The various procedures are outlined in what follows.

3.8.1 Score-distribution parameters. Total score distribution parameters (means and variances) were estimated in the usual way for the experimental and each of the control groups. Since each of the distributions showed symmetry, t-tests of significance of differences between means were performed. Specifically, comparisons were made between E and Cc; E and Ctf; Cc and Ctf; E and Cy; E and Cp.

3.8.2 Reliability (homogeneity). Reliability is an index reflecting the proportion of error-variance in total variance among a set of test scores. It is commonly expressed in the general form of a correlation coefficient and is therefore constrained to a maximum value of +1.

One source of error variance on a test results from inconsistency of performance of testees on test items. One approach to testing is to arrange items in order of difficulty so that at the point where a

testee cannot respond to items, one may assume that he has reached the limit of his capacity. If this break in response-pattern were to occur at different points for different individuals, the test would be said to be perfectly reliable (homogeneous). Less than perfect reliability (a coefficient less than +1) would indicate that a test is not a pure measure of a trait. Low homogeneity coefficients are sometimes criticized as a form of unreliability since they result from tests measuring a number of traits simultaneously. In a number of circumstances such criticism is justified, yet if the purpose of testing is to obtain predictively valid measures, there is both theoretical and empirical basis for having a proportion of heterogeneity in the test.

In this study, the coefficient-alpha modification of the Kuder-Richardson 20 homogeneity formula was used to accommodate the continuous distribution on each item score that results from use of the RSS. The coefficient was calculated on the basis of the formula

$$\alpha_{20} = \frac{N (\sigma_x^2 - \sum pq)}{(N - 1)\sigma_x^2}$$

- where
- N = number of items on the test
 - σ_x^2 = variance of the total observed scores
 - p = proportion of people with item i correct
 - q = 1 - p

3.8.3 Reliability (equivalence). In order to eliminate conceptual problems of equating error variance with total variance in test scores, it has become customary to devise and administer parallel forms of a test. An approximation to creating parallel forms is to divide the items of an administered test into two equivalent parts and to correlate the scores on each part, thus arriving at a measure of equivalence. The reliability of the test as a whole can be estimated from this coefficient by subsequent application of the Spearman-Brown prophecy formula.

A problem remains in equating the two portions of the test. At least three practical solutions have been advanced. One such method is to divide the test by placing odd-numbered items in one grouping and even-numbered items in another. A second alternative, and one that gives a lower-bound estimate of reliability, is to take a random split of items. A third and more rigorous alternative is to match items on the basis of difficulty, discriminability and content. The coefficient resulting from this matching gives an upper-bound estimate of reliability. In this instance, items were matched according to difficulty, this being the only available criterion since items had been deliberately constructed to differ on the other criteria. Both upper and lower bound estimates of reliability were calculated.

3.8.4 Test-criterion correlation ("validity"). The question of establishing a criterion was essentially one of arbitrary choice. The point was that confidence-testing provides a different quality of information from any other source so that it was a priori unlikely to provide any impressively high validity coefficient with available measures. Therefore, it was simply a matter of choosing some set of data that would give a comparison. It so happened that students in the experimental and control groups had recently completed their school term-examinations. As merely a gross measure of some kind of relation with an external set of data the test scores were correlated (product moment correlation, r) with the examination results. Three correlation coefficients were obtained: correlation between school examination scores and test responses scored according to RSS; school examination scores and the same test responses scored conventionally; and between conventional and RSS scores.

3.8.5 Item-test intercorrelation. Under circumstances in which an item-response is scored either zero or one, the most appropriate correlation coefficient is the point biserial coefficient, which is an approximation to product-moment correlation (r) when adjustment has been made for the dichotomy. A technical advantage to confidence-testing is that the items yield a score-continuum. Therefore, no approximation to r is required. As a measure of item-test homogeneity (reliability), the product-moment coefficient was calculated for each of the 56 items.

The standard error of the mean (SE_m) was calculated for each item, where

$$SE_m = [\sqrt{pq}(1 - r_{xx})^{0.5}]$$

and where r_{xx} is the reliability and pq the item variance. This index helps in the interpretation of reliability.

3.8.6 Item-criterion intercorrelation. A frequently calculated validity index is the correlation of an item score with a criterion score. This kind of analysis indicates the extent to which any given item is contributing to the predictive utility of the test as a whole. In the present circumstances, the problem of finding a suitable criterion was even more potentially spurious than in the case of establishing total-test validity. What would have to be available to provide the desired validity-assessment of confidence-testing would be a set of measures derived from a similarly-constructed and similarly-intentioned instrument. Such criterion measures are just not available and current opinion is that the computation of validity-coefficients for confidence testing items in the absence of such a criterion is essentially a waste of time. These several considerations resulted in the judgment that it was better not to attempt any estimation of item-validity rather than to create a set of uninterpretable coefficients.

3.8.7 Item difficulty. Item difficulty is most simply represented as the proportion of people getting an item correct. Where confidence estimates are involved, an item was judged "correct" whenever the confidence of the testee in the keyed alternative exceeded .50. If two alternatives were keyed as correct, an average confidence in excess of .50 was used to place the item in the "correct" category.

3.8.8 Item discriminability. Discrimination indices provide information as to which items distinguish between those students whose performance on the test places them in the upper 27 per cent of the total score distribution from those whose scores fall in the lower 27 per cent. Although there is a number of ways in which item discriminability may be represented, a common practice is to designate an item as discriminatory if the difference in the proportion of testees in the upper- and lower 27 per cent exceeds ten per cent, when the number passing a given item is compared.

The tests from the experimental and the control groups were scored according to the appropriate procedures. Tests from the experimental group were rank-ordered according to their RSS score and the upper and lower 27 per cent identified. The proportion from each group passing each of the 56 items was found and the discriminability thereby determined.

The purpose of such an analysis is to determine those items which do not contribute to a rank-ordering of testees.

3.8.9 Item characteristic curves. It is sometimes difficult to correlate the meaning or import of item difficulty and discriminability when these indices are presented separately. Item discriminability is, in part, a function of item difficulty. A very difficult item (one which few can answer) is contributing almost nothing to discriminability of a test; similarly an item which is very easy is contributing little. The maximum potential discriminability occurs with an item-difficulty of .5.

Item characteristic curves, which are plots of the total test scores (horizontal axis) against the cumulative proportion of testees passing a given item, provide a visual aid to interpreting the combined information about difficulty and discriminability. An item characteristic curve is sigmoid in form, although because of marginal restraints degenerate curves may result. A curve with a relatively steep slope will indicate a discriminating item; the projection of the point of inflection on the x-axis will indicate the item difficulty.

A number of representative item characteristic curves was plotted.

The next chapter contains the results of these analyses.

CHAPTER 4

PRESENTATION AND INTERPRETATION OF RESULTS

The preceding chapter outlined the various methods that were to be applied to the data. The results of the several analyses are presented in the sections that follow. In order to provide maximum information from the results, necessary commentary and discussion is provided.

4.1 Test parameters and their interpretation.

4.1.1 Test parameters. The results of calculating the mean, variance and standard deviation of test scores are provided in Table 1. The results of testing for the significance of differences between means (t-test) of the experimental and the four control groups are also given; so are the results of testing for the significance of differences between variances (F-test).

The interpretability of these results depends upon their level of significance of difference. The mean score obtained by use of S-RSS is significantly greater at the 1 per cent level than the mean score obtained from the use of the conventional (0,1) scoring system. One may infer from this that conventionally scored tests are not permitting the acknowledgement of partial knowledge which -- in the case of this student population, at least -- was sufficient to produce a significant difference.

As with any such result, one needs to make a distinction between its statistical and its practical significance. The obtained difference, though statistically highly significant, may have less practical import than this difference implies. The actual magnitude of difference was only 3.64 score-points (6.30 percent, if the scores were converted to percentages). A question would have to be raised as to whether the extra effort involved in administering and scoring the test under confidence testing conditions was sufficient to increase the utility of the test. As a matter of fact, a quite markedly different rank-ordering of students results from having administered the confidence-testing procedure (q.v.). Thus, a decision would have to be made not only in terms of utility, but also in terms of the extent to which it was important to make decisions upon the additional information confidence-testing yields. Certainly, the gain in diagnostic information was apparent: this could be sufficient justification in itself for the additional effort required during the testing phase. Again: one would have to evaluate the test strategy on the basis of the purpose for which the results were intended. The evidence indicates that a significant gain does accrue if the effort is warranted in relation to the importance of the decisions to be made.

The other differences which resulted in significant differences between means arose from comparisons between the experimental- and the control group Cp; and between tests' mean scores when different instructions were given. The first of these differences may be interpreted as a clear indication that the test was indeed measuring a state of

TABLE 1

TEST PARAMETERS: SIGNIFICANCES OF DIFFERENCES

Treatment conditions		Mean	Variance	Standard deviation
Experimental: confidence instructions				
S-RSS scoring	<u>E</u>	28.85	19.03	4.36
Control: confidence instructions				
conventional scoring	<u>Cc</u>	25.21	21.10	4.58
Control: conventional instructions				
conventional scoring	<u>Ctf</u>	28.40	28.21	5.33
Control: confidence instructions				
S-RSS scoring	<u>Cy</u>	28.70	14.12	3.76
Control: confidence instructions				
S-RSS scoring	<u>Cp</u>	19.52	12.67	3.56

Significances of differences between means (t-test):

Comparison:	<u>E</u> - <u>Cc</u>	:	3.63**	df = 78	
	<u>E</u> - <u>Ctf</u>	:	.41	df = 95	(NS)
	<u>Cc</u> - <u>Ctf</u>	:	3.16**	df = 95	
	<u>E</u> - <u>Cy</u>	:	.20	df = 58	(NS)
	<u>E</u> - <u>Cp</u>	:	9.21**	df = 61	

Significances of differences between variances (F-test):

Comparisons were made between each of the above variance-pairs. None was significant at the 5% level-of-significance. (F-max. = 1.482, df 57/40).

* significant at 5% level

** significant at 1% level

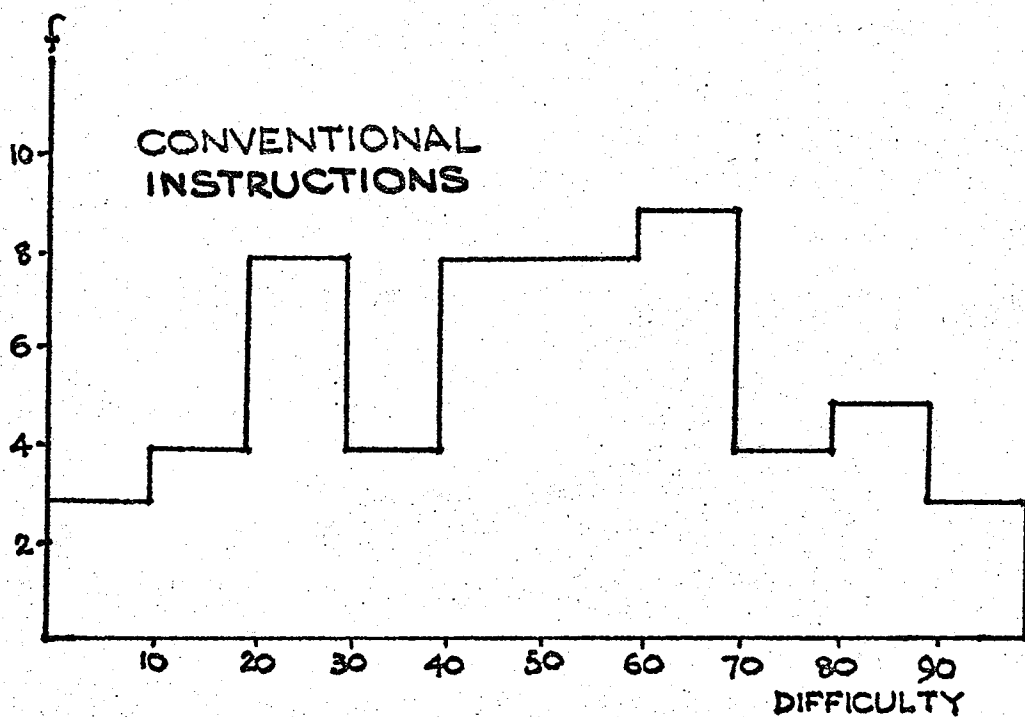
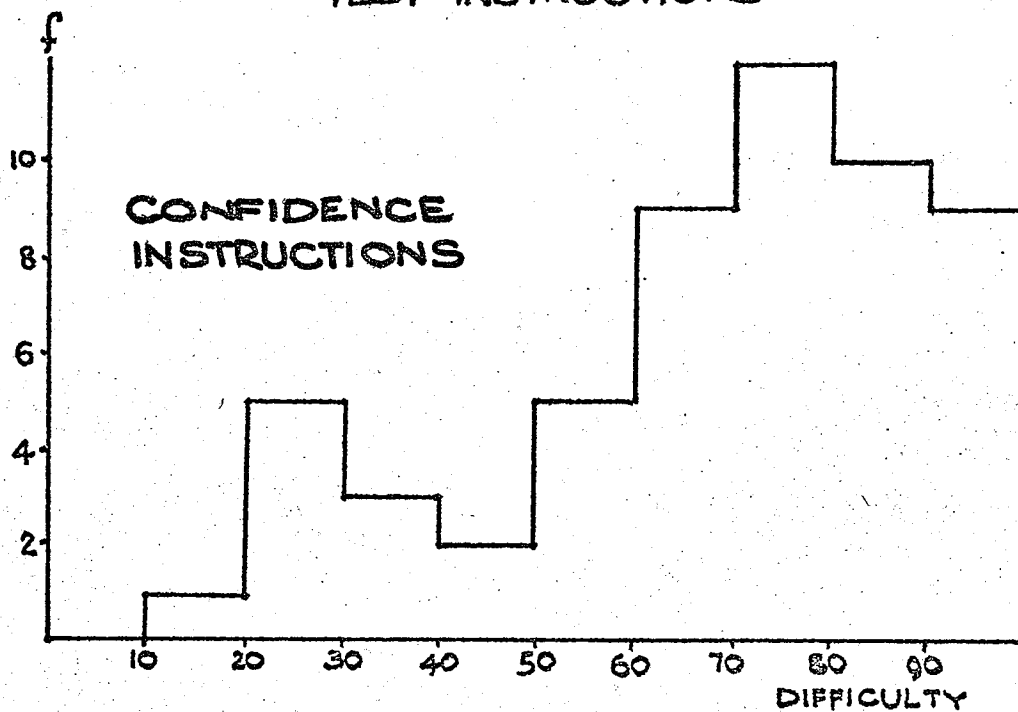
biological knowledge, since the Cp control was comprised of students without any previous formal instruction in biology.

The remaining significant difference resulted from a comparison between test mean-scores when the tests were administered under true-false instructions and under confidence-testing instructions. The outcome implies that testees found it easier to respond in the conventional mode than to respond by indicating their confidence in the correctness of each alternative provided. This general observation was confirmed by considering the item difficulties (q.v. and Figure 1) and by affective responses from the testees (section 4.4). In Figure 1, the negative skew of the item-score distribution under confidence-testing instructions lies in sharp contrast to the relative symmetry of the item-score distribution under conventional instructions.

A somewhat different interpretation of the same outcome resides in the psychological phenomenon known as "response set." "Response set" (Cronbach, 1946) is defined as "any tendency causing a person to give different responses to test items than he would when the same content is presented in different form". The best-known of response sets that have been identified are those representing tendencies to acquiesce and to disagree, regardless; to guess; and to take extreme positions. The mean differences obtained under different instructions may well be an instance of test-taking rigidity in the way responses are to be made.

The fact that the mean difference E-Cy was not different

FIGURE 1: ITEM DIFFICULTY AS A FUNCTION OF TEST INSTRUCTIONS



beyond a chance level means that student performance was independent of a particular approach to biology. (It will be recalled that the E group was taking the Blue Version course; the Cy group was taking the Yellow Version course). The inference was that the items as a whole were measuring general biological principles.

4.1.2 Sex differences. In the 1961-62 BSCS program evaluations, a number of variables was found to be related to differential student performance (Grobman, 1968, p.44). Among the variables that the BSCS evaluators found to be important was the performance difference between boys and girls. There is sufficient data documenting boy-girl differences in performance on a variety of achievement and aptitude measures to warrant attention to this factor. For this reason it was deemed useful to make relevant comparisons between mean performances of boys and of girls. These comparisons are summarized in Table 2 (over).

No significant differences were found between mean scores in any of the comparisons made. In examining the variance in performances, however, a number of differences were brought to light. For boys, there was not a significant difference in variance under the different scoring conditions: for girls, the variances differed significantly at the five per cent level. The S-RSS procedure did not result in a significant difference being obtained when the boys' and the girls' variances were compared, but the conventional procedure resulted in a difference that was significant at the one per cent level.

TABLE 2
 TEST PARAMETERS: SIGNIFICANCES OF DIFFERENCES
 BETWEEN BOYS AND GIRLS (CONFIDENCE INSTRUCTIONS)

Treatment conditions			Mean	Variance	Standard deviation
Boys:	S-RSS scoring	B_S	27.22	659	25.6
Girls:	S-RSS scoring	G_S	28.33	576	24.0
Boys:	Conventional scoring	B_C	22.02	481	21.9
Girls:	Conventional scoring	G_C	23.13	1149	33.9
.					

Significance of differences between means (t -test):

Comparison:	$B_S - B_C$:	.706	df = 50	(NS)
	$G_S - G_C$:	.674		(NS)
	$B_S - G_S$:	.422		(NS)
	$B_C - G_C$:	.398		(NS)
.					

Significance of differences between variances (F-test):

Comparison:	$B_S - B_C$:	1.371	df = 21/21	(NS)
	$G_S - G_C$:	1.994*	df = 29/29	
	$B_S - G_S$:	1.144	df = 21/29	(NS)
	$B_C - G_C$:	2.389**	df = 21/29	
.					

* significant at 5% level

** significant at 1% level

These findings have several implications. For girls, the greater variance in performance when tests were administered according to confidence-testing instructions but were scored conventionally means that these students were, indeed, following instructions as to how to distribute their points. The large variance meant that points were being allocated to all alternatives, and this, in turn, implies a lack of information. The girls were, in other words, accurately reflecting their states-of-knowledge.

The obtained significance of difference in variance when the performance of boys and of girls was compared implied that the boys were performing more homogeneously than the girls were, and were probably not accurately reflecting their states-of-knowledge. The variance obtained under conventional scoring procedures was found to be significant at the one per cent level. This highly significant difference was contrasted to a nonsignificant difference in performance variance under confidence-testing scoring procedures. Therefore, if a choice of scoring were to be made, the choice would inevitably have to fall to S-RSS scoring procedures because these latter do not artificially introduce a sex differential.

4.2 Reliability.

The obtained upper- and lower-bound estimates of reliability were .58 and .55 respectively (Display 2). Test homogeneity was calculated as .50. Each of these coefficients is not high in the

generally-accepted sense of test reliability but is comparable to that found elsewhere under similar conditions (Hambleton, et al., 1970). The reliability of the scores under conventional scoring procedures was even less: .44. The gain of reliability in the confidence-testing procedures may be the result of a specific attempt to design a test that is appropriate to the method.

The test homogeneity indicates the degree to which the test as a whole is a measure of a certain trait as against being a measure of intraindividual differences. The obtained alpha-value suggests that there was some inconsistency in student performance on the items. Also, since it is a logical necessity in test construction that the test content be as heterogeneous as the subject-matter being tested, the calculated value of .50 may be reflecting the heterogeneity of the test content. A test of the general subject-matter area "Biology" could not be expected to be as homogeneous as tests of, say, vocabulary, chemical symbols, addition of integers and the like.

Of theoretical necessity, furthermore, is the fact that the smaller the variance in the experimental group, the smaller the calculated test homogeneity. One might not -- in other words -- anticipate high homogeneity-indices from tests administered to homogeneous groups of students. The classes in the experiment were, as it turned out, relatively homogeneous in ability and this external fact probably influenced the magnitude of the coefficient.

DISPLAY 2

TEST RELIABILITY AND VALIDITY

Reliability (Equivalence)

A random split of the test into two parallel halves resulted in the lower-bound estimate of reliability.

$$\text{Corrected } r = .55$$

A matched (difficulty) split of the test into halves resulted in the upper-bound estimate of reliability.

$$\text{Corrected } r = .58$$

Homogeneity

Homogeneity was calculated by the alpha-20 coefficient.

$$\alpha_{20} = .50$$

Validity

Scores on school biology examinations were arbitrarily selected as a criterion for estimation of empirical validity. Three coefficients of correlation were calculated, where, x = criterion score; y = S-RSS score; z = conventional score, confidence instructions.

$$r_{xy} = .41$$

$$r_{xz} = .72$$

$$r_{yz} = .68$$

4.3 Validity.

Sometimes -- and fairly -- a criticism is made of test constructors who fail to report validity because of a lack of suitable criteria. The claim is made that constructors are simply avoiding what is probably the most rigorous of tests. In order to circumvent such criticism, the arbitrary standard of school examination results was used here as criterion. The validity coefficients that resulted (see Display 2) ranged from .4 to .7. These outcomes are not surprising. They may best be interpreted as indicating that the rank-ordering of testees would change considerably if a change were made from conventional to confidence-testing procedures. This result confirms previous findings (Shuford et al., 1966).

Concurrent validity coefficients such as these, generally are low estimates of true validity. Furthermore, the diagnostic goal that is subscribed to by confidence testing does not require validity beyond the immediate situation in order for the test to be most useful in a diagnostic sense. The diagnostic utility of such a method for formative, decisions does not require high predictive or construct validities.

4.4 Affective response to confidence testing.

Responses to the questionnaire distributed for the purpose of assessing affective reaction to confidence-testing may be considered a form of cross-validation. The latter is a process of gathering new information on test-effectiveness subsequent to the utilization of test

scores. Failure to obtain such data may lead to exaggerated claims of test-effectiveness.

In response to questions about length of the test, 40 per cent of the testees considered that the instrument was too long; 50 per cent that it was about right. Sixty-five per cent found that five alternatives per item was too many; 33 per cent that the number was about right; and three per cent would have preferred more alternatives.

The reading-level of the items was judged to be about right by 67 per cent of the respondents, while 33 per cent found it difficult. The group was about evenly split in their opinion of the amount of memory required: 48 per cent believed a lot of memory-work was needed; 52 per cent replied that the amount of memorization required was normal.

Confirming the objective data, 68 per cent of the testees indicated that they could not have reasoned the answers on the basis of general knowledge -- that they did, indeed, need specific biological knowledge. The remainder believed that they could have guessed at more than a few of the items without biological "expertise".

Eighty per cent of the testees found that distributing ten points was a difficult task in comparison to marking just one alternative correct; five per cent found it easier and the remainder found that there was no difference in the difficulty of responding to items. Eighty nine percent claimed that they guessed more often under confidence testing conditions; to the remainder, the instructions made no difference.

Of course, what would be labelled as "guessing" by a student would be called "state-of-knowledge" by a teacher using confidence-testing methods.

The next response was unanticipated. In reply to a question, "Does distributing ten points indicate just how much you know....", 40 per cent indicated that the point-distribution method showed they knew less; 36 per cent that it made no difference; and 24 per cent that it showed they knew more! This finding implies that students perhaps did not understand the underlying purposes of confidence-testing as much as had been thought. In sum, 82 per cent of the testees expressed a preference for marking one alternative correct, only; six per cent preferred confidence-testing and the remaining 12 per cent were indifferent as to method.

The overwhelming majority -- 92 per cent -- believed that the possible multiplicity of response made the items more difficult. Fifty four per cent indicated that they realized that an equal distribution of two points per alternative resulted in a higher score than guessing blindly and putting all ten points on one (wrong) alternative; 46 per cent claimed that they did not realize this, again suggesting that the testees did not understand that responding honestly to each item would maximize their scores.

These affective responses confirmed the numerical findings in all instances and in addition cast some light onto possible obscurity

that might exist in regard to the confidence-testing procedure as a whole. In turn, these misunderstandings may have contributed to a lowering of the obtained reliability and validity coefficients.

4.5 Item analysis.

Results from carrying out the various item analyses are provided in Tables 3 and 4 and in Figure 2.

4.5.1 Item difficulty. Item difficulties (p) and the item difficulty index (Q); item means (p) and variances (pq); and that proportion of testees passing the item in the upper and lower 27 per cent of respondents are each presented in Table 3.

On the whole, the items may be considered as difficult: for 46 of the 56 items, fewer than half the respondents got the item correct. If one regards the item-set as constituting a test the range of item difficulty was from $p = .02$ to $p = .86$, thereby probably lowering the total test reliability. A priori reliability is maximized for a test made up of items each of which has a $p = .50$. The obtained range is considerable by any standards.

Reference back to Figure 1 draws attention to the relative difficulty of items that were responded to under confidence-testing instructions as compared with conventional true-false instructions. The difference in the symmetry suggests that the testees apparently found the more complex instructions increased overall item-difficulty.

TABLE 3
ITEM CHARACTERISTICS

Item	U	L	M	T	p	pq	Q	Item	U	L	M	T	p	pq	Q
1*	33	18	33	84	42	244	58	29*	19	04	12	35	17	141	83
2	18	19	27	64	32	219	68	30*	09	03	09	21	10	090	90
3*	49	36	60	145	72	202	28	31*	25	10	45	80	40	240	60
4*	37	23	62	122	61	234	39	32*	17	04	15	36	18	148	82
5*	24	15	42	81	40	240	60	33*	13	04	16	33	17	141	83
6	09	07	20	36	18	148	82	34*	35	21	51	107	53	250	47
7*	21	03	09	33	16	134	84	35*	26	17	33	76	38	236	62
8*	28	18	46	92	46	248	54	36*	41	34	65	140	70	210	30
9*	31	18	42	91	45	248	55	37*	13	14	27	54	27	197	73
10*	17	05	12	34	17	141	83	38	07	06	13	26	13	113	87
11*	30	20	42	92	46	248	54	39*	42	25	64	131	65	228	35
12*	46	37	68	151	75	188	25	40*	49	37	75	161	80	160	20
13*	47	36	75	158	79	166	21	41*	32	16	42	90	45	248	55
14*	21	10	28	59	29	206	71	42*	23	15	37	75	37	233	63
15*	23	08	20	51	25	188	75	43*	24	11	26	61	30	210	70
16	37	32	80	149	75	188	25	44*	22	05	28	55	27	197	73
17*	36	12	32	80	40	240	60	45*	23	11	24	58	29	206	71
18*	13	04	20	37	18	148	82	46*	09	00	08	17	08	074	92
19*	35	09	33	77	38	236	62	47	02	00	06	08	04	038	96
20	02	03	03	08	04	038	96	48	05	00	01	06	03	029	97
21*	53	41	79	173	86	120	14	49*	27	09	38	74	37	233	63
22	29	29	45	103	51	250	49	50*	07	00	09	16	08	074	92
23	01	01	02	04	02	020	98	51*	17	08	18	43	21	166	79
24*	11	03	16	30	15	128	85	52*	19	05	19	43	21	166	79
25*	26	09	19	54	27	197	73	53*	06	00	06	12	06	056	94
26*	22	09	21	52	26	192	74	54*	17	12	22	51	25	188	75
27*	13	06	18	37	18	148	82	55	06	04	06	16	08	074	92
28*	23	09	29	61	30	210	70	56*	19	13	33	65	32	218	68

Explanation of symbols:

* indicates a discriminating item (U - L greater than 10%)

U number passing item of upper 27% total test scores

L number passing item of lower 27% total test scores

M number passing item of middle 46% total test scores

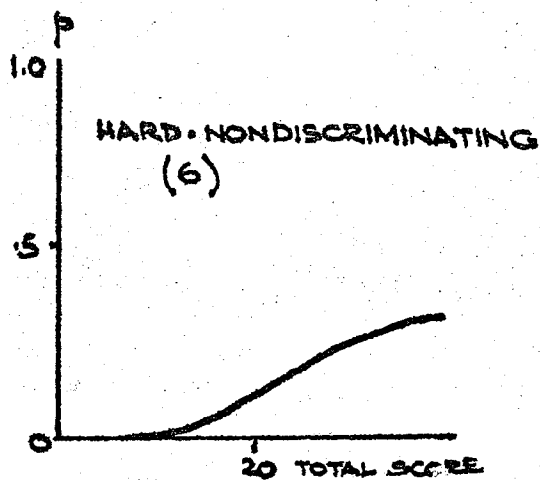
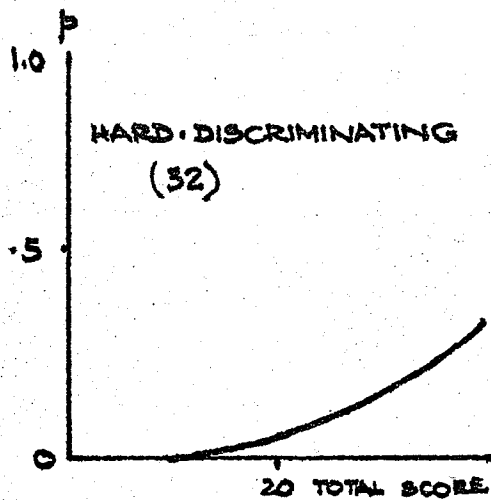
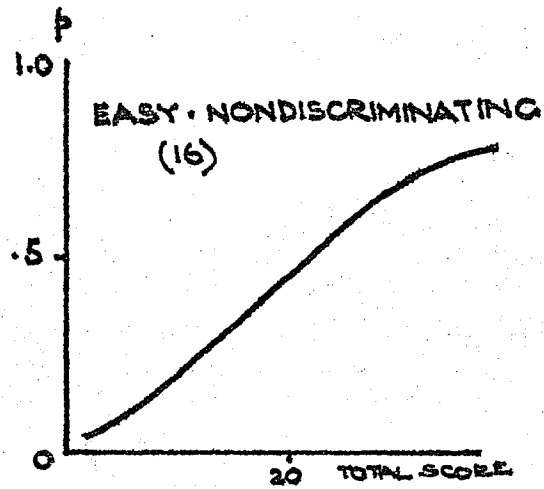
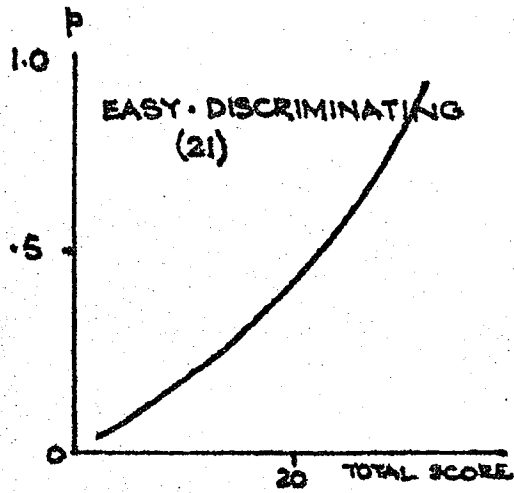
T total number passing item (U + L + M)

p proportion passing item (decimal omitted)

pq item variance (decimal omitted)

Q difficulty index, percent not passing item

FIGURE 2: ITEM CHARACTERISTIC CURVES



4.5.2 Item discriminability. Item discriminability may also be inferred from Table 3. Those items marked with a star (*) are discriminating in the sense that different proportions of testees in the upper- and the lower 27 per cent got the item correct. The conventional difference of ten per cent was adopted in determining discriminability.

Although the items were difficult, only ten of them proved to be nondiscriminating. Lack of discriminability may be construed as a criticism of an item only if the item-content is judged not to be important enough for inclusion in a test on content-validity grounds.

4.5.3 Item-characteristic curves. Four item-characteristic curves are provided in Figure 2. These items were chosen because they respectively represented the logical combinations of difficulty and discriminability. General differences in the shapes of the curves are apparent.

4.4.4 Item-test intercorrelations. Item-test intercorrelation and item standard error of measurement is provided for each of the 56 items (Table 4). A small standard error of measurement is interpreted as indicating good measurement, whereas a large standard error may be interpreted as either that the items are poor or that the group of testees is of less than high-ability. Since the test is known to be difficult, yet the standard error of measurement is large, and from external evidence from the schools of the testees, one must conclude

TABLE 4
ITEM-TEST INTERCORRELATIONS AND
STANDARD ERRORS OF MEASUREMENT

Item	r	SE _M	Item	r	SE _M
1.	-.17	.53	29.	.33	.31
2.	.62	.29	30.*	-.13	.32
3.	.13	.42	31.	.08	.47
4.	.30	.41	32.	-.28	.44
5.	.52	.34	33.	-.16	.41
6.	-.10	.40	34.	.44	.37
7.	.40	.28	35.*	.23	.43
8.	.32	.41	36.	.38	.36
9.*	.18	.45	37.	.11	.42
10.*	.38	.30	38.	.11	.32
11.	.40	.39	39.	.41	.37
12.	.60	.27	40.	.21	.36
13.	-.05	.42	41.	.23	.44
14.	.28	.39	42.	-.02	.49
15.	.29	.37	43.	.26	.39
16.	.06	.42	44.*	.13	.41
17.*	.31	.41	45.	.31	.38
18.	.16	.35	46.*	.24	.24
19.*	.33	.40	47.*	.29	.17
20.	.23	.17	48.*	.45	.13
21.	.18	.31	49.	.11	.46
22.	-.51	.61	50.*	-.16	.29
23.*	.69	.08	51.	.39	.32
24.*	.10	.34	52.	.32	.39
25.	.39	.35	53.*	-.29	.27
26.	.44	.33	54.	.06	.42
27.	-.19	.42	55.*	.32	.22
28.	.56	.30	56.	-.33	.54

* Indicates an item keyed with two
correct alternatives.

that the error effect is due largely to a lack of ability in the testees. The more general point is that one cannot conclude that confidence testing procedures are invalid simply because standard errors of measurement are large.

It was of further interest to note that those items that had more than one alternative keyed as correct did not prove to be consistently different in item-test intercorrelation. The belief of the students that these items were more difficult was also not verified.

These results lead to a number of conclusions; to the identification of a number of areas in which further research is needed and to some general evaluative comments about confidence testing. Some of these summary statements are ventured in the final chapter.

CHAPTER 5

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary.

With the use of standardized and teacher-made tests having become established practice in today's schools, it was a purpose of this study to investigate a method of extracting from objective tests more information concerning a student's state of knowledge than is possible from a conventional testing program.

Multiple-choice tests scored conventionally (0, 1) serve to separate the informed testee from the misinformed, but do not allow for the identification of partially informed and uninformed testees. Further, the test-wise student will realize that his best test-taking strategy, under the conventional scoring system, is to guess if he is unsure of an answer, so that with luck he may be classified as -- and indistinguishable from -- the truly informed testee. Moreover, conventional testing methods by requiring a student to place all his confidence in one, and only one, alternative, are contributing to a disparity between philosophy and practice by implying that every problem has a single, immutable, correct answer.

The very process of testing is justified only insofar as it contributes an increased utility over base-rate decisions. In terms of current preoccupation of society with arranging people in hierarchical

orders for some specific purpose, it is of paramount importance to realize that not all testing procedures are convergent in their rank-ordering of individuals. Any given sequencing of ability and aptitude can be considered -- at least in part -- a function of the choice of test instrument and scoring procedure. It has become critical for society that, in the face of a widening technology and increasingly flexible value-systems, fine discriminations can be made in the human resource pool. Whereas only a short time ago it was sufficient for individuals to acquire relatively general skills and concepts or ideas in order to be able to function, today not only must people have these general skills to withstand the threat of obsolescence but also society demands the identification of special talent in order that individual skills will contribute maximally to the common weal. Therefore, the onus is upon the psychometrician to devise testing strategies that will permit those discriminations to be made that will capitalize upon the exact state of knowledge of every individual.

During the past decade, workable techniques for the extraction of maximum information from test items has been proposed by a number of authors. These techniques operate on the possibility of responding to more than one alternative of a multiple-choice item by ranking the alternatives or by assigning a number reflecting degree-of-confidence in the correctness of each alternative. Such systems as the latter include the admissible category system, which involves more than one response category per alternative; the confidence-weighting system; the

differential-weighting response-alternatives method; and admissible probability measurement procedures. These various techniques are referred to collectively as "confidence testing" procedures and stipulate methods for finding the subjective probability of respondents.

Historically, the concept of "degree-of-belief" was identified early in the eighteenth century. Another concept, concerning the worth of outcomes ("utility") arose from concerns in economics. The conjoining of utility theory and what is now known as subjective-probability theory into the more general decision theory, permits the description of all behavior in which there is mathematical, if not moral, consistency. Decision theory has recently been used to formulate the quality of institutional decisions made on the basis of test information. Static, risky, decision-making situations (such as testing is) rely upon subjective estimates of both utility and probability.

Although it has been suggested that confidence-testing procedures could greatly increase the amount of information available from a test, experimental attempts to measure confidence have generally failed, owing in large part to the inadequate structure of the scoring system. Decision theory has been used to develop procedures for measuring an individual's confidence, or subjective probability. Several scoring systems have been proposed, each system having the property that an individual could maximize his expected utility (score) if, and only if, he honestly expresses his subjective probability (confidence) about the correctness of each alternative.

Claims have been made that responses under confidence-testing instructions are more valid and more reliable than under conventional test-taking instructions. Confidence-testing scores result in a different ranking of individuals from the ranking obtained from conventional scoring and also provide a higher total score since credit is given for part knowledge. On these grounds it has been argued that conventional testing underestimates the achievement of many individuals and thus contributes to incorrect decisions. Further, since confidence scoring places item-scores on a continuum instead of a 0-1 dichotomy, there is the possibility for finer discrimination. This is particularly valuable for diagnostic assessments, formative evaluation and item-writing. Various response aids have been developed for student self-avaluation using confidence testing.

The effects of guessing has been a constant concern in the theory of testing. Under conventional testing conditions, the problem of correcting for guessing is largely ignored because of the difficulty of detecting and preventing it. Guessing has been variously defined as (i) not answering a question on the basis of surety about the correct response but on the basis of some moderate surety ("rational guessing"); (ii) answering a question when all possible alternatives are considered equally likely ("blind guessing" corresponding to a state of being uninformed): (iii) answering a question when the alternative chosen is regarded as being equally likely with some, but not all of the answers ("partially blind" guessing, corresponding to a state of misinformation).

The only scoring system which can distinguish all three types of guessing are admissible-probability measurement procedures. Conventional scoring cannot distinguish any of the three and no discrete choice system can identify (i) and (ii) reliably. Neither of the latter systems can provide a method for identifying (iii).

Confidence testing provides a technique by means of which a testee may maximize his score without resorting to guessing. By distributing his points honestly, a testee will be assured of a score reflecting his state of knowledge. Under the conditions that would encourage guessing in a conventional scoring situation (when the student is unsure of any or all of the alternatives), by distributing his points equally across the attractive alternatives the testee will be guaranteed a score that is greater than the chance level. Conventional scoring permits the situation where two students with the same quantum of knowledge may, by capitalizing differently on chance, achieve two widely different scores. Given that students respond accurately, confidence testing precludes this possibility.

This particular study had three objectives: (i) to provide a bridge between an agreed-upon philosophy of science and a means of testing that would realize that philosophy, that is, realize the tenet of uncertainty; (ii) to provide a means of testing which eliminates the need for guessing and by capturing accurate responses increases the utility of a test; and (iii) to employ a scoring system which credits part knowledge.

Three hundred students were tested. The sample was divided into an experimental and four control groups. The experimental group was tested under full confidence-testing conditions; the controls were tested, respectively, under conditions of confidence-testing instructions with conventional scoring; conventional instructions with conventional scoring; different content-background in biology; and lack of biology background.

A set of 56 items was selected from a pool of over 200 items written specially for the confidence-testing format. The content of the items was BSCS Blue Version Biology which happened to be undergoing trial in a number of local schools at the time. The 56 items were selected on a number of acknowledged a priori criteria and assembled into standard test format. The test was administered under the appropriate treatment conditions to the students.

The student responses were scored and test parameters found. Certain basic comparisons were made between means and between variances that resulted from the several treatment conditions. It was found that significant differences resulted between the experimental group and the control for scoring and instructions and the control for non-specific biology content, indicating that -- respectively -- confidence-testing procedures produced higher mean scores by giving credit for part knowledge; that testees found conventional instructions easier than confidence testing instructions; and that the items were indeed testing for biology content, though not specifically that of the Blue Version.

No significant differences were found between the mean scores of boys and of girls, but in comparing the variances, significant differences were found between the girls' scores under different scoring conditions and between the girls' and boys' scores under conventional scoring conditions. These differences were attributed to the girls' performances being more spread-out by the confidence scoring procedures than were the boys'. This implied that the girls were responding more in accord with the instructions than were the boys. The sex differential was eliminated by employing confidence scoring.

Test and item-test reliabilities were calculated. The several estimates of reliability were of the order of .55 -- low, but in line with previously-found reliability coefficients for confidence-testing. Reliability under conventional scoring was .44. Item-test reliabilities ranged from -.5 to .7. Standard errors of measurement were relatively large, but both these results were attributed to the uniformly medium ability-level of the group.

Test validity was computed against an arbitrary criterion of school examination scores. Validity for the confidence testing procedure was .4; for the conventionally-scored test, .7. Neither of these results is impressive, but both can be explained in terms of the criterion that was selected, the homogeneity of the group and the difficulty of the items.

Item difficulty and discriminability were also found. On the

whole the items were both difficult and discriminating: 46 of the 56 items were responded to incorrectly by more than 50 per cent of the testees. The items for which there were more than one correct answer were not shown to be more difficult than the one-option items. Only ten of the items were non-discriminating. Item characteristic curves were produced for representative items.

Affective responses were gathered from the students regarding their appreciation of the confidence-testing procedure. These responses confirmed the numerical findings and in addition, indicated that the students were perhaps not totally clear as to the import of following the confidence-testing instructions.

5.2 Conclusions.

Probably the most important evidence for the utility of any test lies in the data concerning its reliability and validity. The few experiments that have been done previously with tests constructed for use in the conventional manner have found reduced reliability and increased validity when corrections for guessing have been applied. Similar results have been obtained when a conventionally constructed test has been scored in a confidence-testing manner. Here, it was found that with a specially-constructed test the reliability increased, though validity was low owing to a lack of a logically parallel criterion. It may be concluded, therefore, that specially-constructed tests will, in fact, increase reliability and since the principal utility of confidence

testing lies in its diagnostic output, one might suggest that the overall payoff from testing is greater under confidence-testing methods. Until criterion measures are available upon the same subjective-probability scale, one would have to withhold judgement as to the appropriateness of confidence-testing for obtaining valid measurements.

Disadvantages that were unearthed were concerned primarily with complexities in comprehending the instructions and the response mode. It was apparent that more time explaining the procedures and more practice in their application would have resolved some of the confusion of the testees. Indubitably, some of their difficulty may be attributed to a reluctance to shift the response-mode, a conclusion affirmed by the essence of the affective responses. Some of the difficulty may also be attributed to a failure of the students to prepare the content of the test adequately thus increasing a priori the test difficulty.

On the positive side, a number of advantages were discovered. The students, almost without fail, scored higher when scored by confidence-testing procedures. These higher scores resulted from part knowledge being credited; this information would have been lost under conventional scoring systems. If completeness of information is a goal of testing, then the complexity of scoring is more than offset by the increase in information from confidence-testing and the therefore increased diagnostic utility of the test. Further, it may be assumed that guessing is less likely to occur -- and even less guessing would occur as the confidence-testing procedure is understood.

Finally, the sex differential artificially introduced in the scores by conventional scoring procedures was eliminated by employment of the confidence-testing method.

5.3 Recommendations.

As a result of having undertaken this study, a number of recommendations emerged:

- (1) confidence-testing, because of its practical utility and diagnostic value, merits further study;
- (2) practical work in the construction of items appropriate to confidence testing must be undertaken to identify optimal strategies;
- (3) such items must be administered to experimental groups in order to accumulate practical experience with these items;
- (4) effort must be made to mechanize the scoring processes so they are applicable by any evaluator, specially in large-group situations, where large numbers of items are involved, and where frequent testing is undertaken;
- (5) current theories of reliability, validity and item-analysis must be examined in the light of confidence-scoring procedures;
- (6) evidence must be gathered concerning the relative impact on a student of receiving a zero item-score or of openly declaring he is uninformed.

APPENDIX

TABLE OF ITEM SCORES (S-RSS FORMULA)

TO BE ALLOCATED TO ALL POSSIBLE

DISTRIBUTIONS OF TEN POINTS

r_h	$r_j - r_h$	F	r_h	$r_j - r_h$	F	r_h	$r_j - r_h$	F
10		1.00	4	321	.73	2	422	.38
9	1	.99	4	33	.69	2	3311	.41
8	2	.98	4	42	.67	2	332	.39
8	11	.98	3	7	.39	2	3221	.43
7	3	.92	3	61	.44	2	2222	.45
7	21	.95	3	511	.50	1	9	.11
7	111	.97	3	52	.49	1	81	.12
6	4	.83	3	4111	.57	1	72	.14
6	31	.88	3	421	.55	1	711	.14
6	211	.92	3	43	.51	1	63	.15
6	1111	.95	3	331	.57	1	621	.15
6	22	.91	3	3211	.61	1	6111	.16
5	5	.71	3	2221	.64	1	54	.15
5	41	.77	3	322	.59	1	531	.17
5	311	.83	2	8	.24	1	522	.17
5	2111	.88	2	71	.27	1	5211	.18
5	221	.86	2	62	.30	1	441	.17
5	32	.81	2	611	.31	1	432	.18
4	6	.56	2	53	.32	1	4311	.19
4	51	.62	2	521	.34	1	4221	.20
4	411	.69	2	5111	.35	1	3321	.20
4	3111	.76	2	44	.33	1	333	.19
4	2211	.78	2	431	.36	1	3222	.21
4	222	.76	2	4211	.39	0		0

CONFIDENCE INSTRUCTIONS

INSTRUCTIONS

*Please read the following instructions carefully, as they are explained:

First, on the ANSWER SHEET that accompanies this test, put your name, school and grade in the spaces provided;

The items are ordinary multiple-choice, but this test is somewhat different from the usual kind of test:

Any particular item may have MORE THAN ONE CORRECT ANSWER. Instead of marking just one alternative as "best", you are to make a response to each alternative. Now, not all answers may be equally good, so we ask you to show us that you know this, by asking you to distribute TEN POINTS among the alternatives. The solution you think is best, you will give most points to; that which is next best, next most points to, and so on; any alternative you consider wrong you would give zero points.

Thus, the second way in which this test differs from the usual kind of test is that you will DISTRIBUTE TEN POINTS AMONG THE ALTERNATIVES, ACCORDING TO YOUR CONFIDENCE IN WHICH ONES ARE CORRECT.

In reading the items, make sure you notice when you are required to respond to a "not" or "except" type of question.

Now, try these two EXAMPLES for yourself.

1. Evidence which led to Watson's and Crick's DNA model was obtained from

(a) X-ray diffraction photos of DNA	a	b	c	d	e
(b) chemical analyses of DNA					
(c) studies of pneumococcus					
(d) studies of bacteriophages					
(e) electron microscopy					
	-----	-----	-----	-----	-----

2. Which of the following is not found after hydrolysis of DNA?

(a) adenine	
(b) guanine	
(c) thymine	
(d) cytosine	
(e) uracil	

The keyed answers to these questions are

1.	5	5	0	0	0
	-----	-----	-----	-----	-----
2.	0	0	0	0	10
	-----	-----	-----	-----	-----

Are there any further questions?