DOCUMENT RESUME

ED 091 427 TM 003 641

AUTHOR Stegman, Charles E.
TITLE Subjective Probability and the Administration of Objective Tests.
SPONS AGENCY Pittsburgh Univ., Pa. School of Education.
PUB DATE [73]
NOTE 37p.; Paper presented at the Annual Convocation of the Northeastern Educational Research Association (4th, Ellenville, New York, October 31-November 2, 1973)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS Annotated Bibliographies; *Confidence Testing; Guessing (Tests); *Measurement Techniques; Multiple Choice Tests; *Objective Tests; Research Needs; Scoring; Testing; Test Reliability; Test Validity

ABSTRACT

Probabilistic testing involves having the examinee assign probabilities to each of the options of a multiple-choice item. These probabilities reflect the student's perception of the correctness of each option. What is presented in the paper is a rationale for probability testing, the current theoretical and empirical findings, and some suggested directions for further research. The rationale given for considering probabilistic testing includes the following points. First, testing involves making decisions under uncertainty as do many situations faced every day and as such should be solved by using a subjective probability decision theoretic paradigm. Second, using multiple-choice testing situations may be a good way of teaching the subjective probability decision theoretic paradigm. Third, probability testing procedures should lead to more reliable and possibly more valid tests. Fourth, probability testing in conjunction with specific utility functions yields a way of incorporating and handling "risk" and "guessing" behavior in testing situations. An annotated bibliography is also included to introduce potential researchers to the general area of confidence testing. (Author)

# Subjective Probability and the Administration of Objective Tests

Charles E. Stegman
University of Pittsburgh

Paper Presented at the 4th Annual Convocation of the
Northeastern Educational Research Association
Fallsview Hotel Ellenville, New York

October 31 - November 2, 1973

# Subjective Probability and the Administration of Objective Tests[1]

Charles E. Stegman
University of Pittsburgh

## Introduction

The widespread use of objective tests began about forty years ago. Two persistent concerns of measurement specialists regarding objective tests since then have been the development of methods for controlling guessing behavior and of taking into account partial knowledge. Failure to take into account guessing behavior and partial knowledge, in the usual 1-0 scoring rule for correct-incorrect responses, has led many to conclude that objective item scores result in a rather crude approximation of a person's actual position on the continuum of the variable being measured.

It may be argued, of course, that this concern is misplaced. If one assumes a homogeneous item set, where the probability of a correct response by person i remains constant across all k items in the set, then one should be concerned with $p_i k$, and not individual item scores, which cannot equal $p_i$ except when $p_i$ is 1 or 0.

While the above is theoretically true, it is also true that decisions are made on the basis of either item scores or small subsets of item scores, that is, subtests or scales of test batteries. The trend toward criterion-referenced measurement indicates that more, rather than less, emphasis will be placed on the evaluation of item responses, where these responses are assumed to represent a sample of behavior(s) from some domain. The homogeneity charac-

[1]This research was supported in part by a grant from the Faculty Research Fund, School of Education, University of Pittsburgh.

teristic is a thorny problem, since homogeneity can be only partially attained. Thus far, the complexity of cognitive processes has kept ahead of the itemwriter's attempts at developing the statistically and psychologically homogeneous item set.

Dissatisfaction with both the conventional methods of administering and scoring objective tests and with the methods advanced to compensate for the various deficiencies has led several measurement specialists to suggest alternative methods for administering and scoring objective tests. These proposed methods have had the common objective of improved precision in the form of greater reliability and validity.

The proposed methods include confidence-weighting (Hevner, 1932; Soderquist, 1936; Ebel, 1965a, 1965b), option-elimination (Coombs, 1953; Coombs et al., 1956), and probabilistic testing (de Finetti, 1965; Shuford et al., 1966). In confidence-weighting the examinee selects the perceived correct option to a multiple-choice question, then indicates his certainty of its correctness on an accompanying confidence scale. Item scores are dependent upon these two factors, accuracy and expressed confidence. Option-elimination requires the student to identify as many of the n-1 distractors as possible from the set of n options for the multiple-choice item. Item scores are a function of the number of correct identifications with a penalty for misidentifying the correct answer as a distractor. Probabilistic testing involves having the examinee assign probabilities to each of the n options of a multiple-choice item. These probabilities reflect the student's perception of the correctness of each option.

In comparing the above three methods it can be argued that confidence-weighting and option-elimination are approximations to probabilistic testing. Confidence-weighting is simply a partial version of the last approach since only one option (the perceived correct answer) is weighted. Also, the confidence

weight assigned is usually limited to only four values, while in probabilistic testing the probability weight can be essentially any number between zero and one. Option-elimination approximates probabilistic testing since the student implicitly weights the options and then dichotomizes the options into two sets: (1) perceived distractors and (2) one or more options thought to contain the right answer. There is no attempt at explicitly measuring the weights attached to the elements of set (1) or, more importantly, set (2) when it contains more than one option. de Finetti (1965), in discussing option elimination, derives formulae for the threshold values necessary for the individual to eliminate a given option. That is, one can work backwards from the options eliminated to set bounds on the probability of "correctness" associated with them.

Echternacht (1972) discussed these three methods under the general heading of confidence testing. The purpose of his paper is "to describe the various forms of confidence testing as they have been developed and to provide a brief evaluation of these forms" (p. 217). His paper presents a good review of literature and overview of the area of "confidence testing."

The present study will limit itself to Echternacht's subcategory "probability testing" which is associated with the personal probability approach of de Finetti. What will be attempted here is to present the rationale for probability testing, to identify the theoretical and empirical findings, and to suggest some directions for further research. It is assumed that the reader is basically familiar with what "probability testing" is, at least at the level of Echternacht (1972), and Lord and Novick (1968, pp. 314-323).

## Rationale

Before considering the "measurement" rationale for probability testing it is important to note de Finetti in 1965 was attempting to *apply* a philosophy of rational decision making under uncertainty to *some* problems associated with

objective testing. This philosophy, which is dependent on subjective probability, is intended to apply to all situations in life involving decision making under uncertainty and taking "objective tests" is only one such situation. A basic postulate of this philosophy is that:

> We are always living and dealing in conditions of uncertainty. If probabilistic thinking is to be the guide in facing uncertainty, it is essential that we learn how to do it 'correctly.' To know the rules of probability and to be acquainted with their practical application is to free us from the danger of inconsistency. (de Finetti, 1970, p. 38)

This postulate is certainly not limited only to subjective probabilists. It has been a basic tenet of psychology for at least thirty years. Egon Brunswik was one of the first psychologists to argue for considering the probabilistic nature of life in designing psychological experiments. In 1943 he stated his position as follows:

> On the whole, only scattered recognition has been given to the fact that object-cue and means-end relationships do not hold with the certainty obtained in the nomothetic study of the so-called laws of nature, but are rather of the character of probability relationships. This deficiency is more clearly reflected in the psychology of learning which has proceeded almost exclusively along a dialectically dichotomized all-or-none pattern of "correct vs. incorrect," "right vs. wrong." Situations in which food can be found always to the right and never to the left, or always behind a black door and never behind a white one, are not representative of the structure of the environment... They are thus not sound as experimental devices from the standpoint of a psychology which wishes to learn, above all other things, something about behavior under conditions representative of actual life... I have expanded on this subject to such an extent because I believe that the probability character of the causal (partial cause-and-effect) relationships in the environment calls for a fundamental, all-inclusive shift in our methodological ideology regarding psychology. (p. 260-261)

Further, when confronted with uncertainty

> All a finite, sub-divine individual can do when acting is--to use a term of Reichenbach--to make a posit, or wa-

> ger. The best he can do is to compromise between cues so that his posit approaches the 'best bet' on the basis of all the probabilities, or past relative frequencies, or relevant interrelationships lumped together. (p. 259)

In a similar manner Hilgard (1951) says:

> A great many perceptual experiences can be understood by considering the perceiving person to be a statistical machine capable of quickly estimating probabilities. That is, each of the cues present now is related to many past experiences. Past experiences provide a kind of table of probabilities according to which estimates are made, but the perceiver has to make use at once of the experience tables corresponding to each of the cues, some of which will point in one direction, some in another. (p. 111-112)

Recently the psychologist David Bakan (1967) has attempted "to gain an understanding of the nature of the learning process through the examination of one particular formulation of the nature of the scientific method, the principle of inverse probability" (p. 58).

To see that the theorists of subjective probability intend to develop procedures applicable to uncertainty as encountered in everyday life one need only consider some of their basic writings. Harold Jeffreys in the preface to the first edition of his book *Theory of Probability* (1961) says: "The chief object of this work is to provide a method of drawing inferences from observational data that will be self-consistent and can also be used in practice" (p. ix). Also that "the fundamental problem of scientific progress, and a fundamental one of everyday life, is that of learning from experience" (p. 1). I.J. Good in his book *Probability and the Weighing of Evidence* (1950) states that "the aim of the present work is to provide a consistent theory of probability that is mathematically simple, logically sound and adequate as a basis for scientific induction for statistics, and for ordinary reasoning" (p. 2). Elsewhere, Good (1965) says that: "The difficulties become clear when it is realized that we estimate probabilities every minute of the day, at least im-

plicitly, and that how we do this is unknown" (p. ix) and "Nevertheless, for purposes of making decisions, we do manage to approximate estimates of probabilities. How this is done is an interesting problem in psychology and neurophysiology" (p. 4).

To quote one other source Alberoni (1962) argues that probabilistic thinking "comes into play every time a man finds himself faced with uncertainty and he must take decisions and a stand with respect to the future while basing himself on uncertain or incomplete knowledge" (p. 285). It is this characteristic of answering multiple-choice questions that led de Finetti to propose the use of alternative measurement procedures. That is, in the usual testing situation the student has to make a decision and take a stand [select one option as correct] with respect to the future [his selection will be graded correct or incorrect] and this decision may be made on uncertain or incomplete information [he knows he does not know the correct answer but is "fairly confident" about the correctness of some of the options]. For the person who is "certain" as to the right answer, the problem of uncertainty does not exist and his best response is to indicate that option. de Finetti's (1965) paper is normative in that he is "not interested here in the actual behavior as it results from the habits or other psychological tendencies of different persons, but in analyzing what response is most advantageous in the face of the uncertainty of any given situation" (p. 87).

Winkler and Murphy (1968) in discussing several uses of probability scoring rules for evaluating meteorologists allude to two other reasons for using probabilistic testing. These are (1) to help people become "better" assessors and (2) to evaluate people in a substantive area. The first reason is closely related to de Finetti's philosophy. That is, the multiple-choice testing situation may be a very good situation for teaching people the fundamentals of pro-

babilistic decision making, i.e., the ability to accurately specify probabilities which reflect the person's subjective beliefs. The application of probability testing is confounded if you do not have "good" probability assessors in the normative sense of possessing some expertise in probability assessment.

The second reason (given by Winkler and Murphy (1968)) is one that most many people in educational measurement would consider most important. Can probability testing be used to evaluate people in a substantive context, and if so, do the procedures yield test scores which are more reliable and/or valid than conventional testing procedures? If the procedures do not increase reliability and validity then some will argue why bother to expend the additional time and money to use them. To quote Lord and Novick (1968):

> Thus, at present, the sole recommendation of these new methods is their strong conceptual attractiveness. In evaluating any new response method, it will be necessary to show that it adds more relevant ability variation to the system than error variation, and that any such relative increase in information retrieved is worth the effort... (p. 314)
>
> As with all other mental test theories, validity of this theory must be established by using it to make and verify important predictions. If the theory of personal probability in application to the assessment of partial knowledge suggests certain measurement procedures and related item-scoring and item-weighting formulas that are then empirically established to be valid predictors, then this theory will have been validated for this particular purpose. (p. 315)

Coombs (1953), Coombs et al. (1956) and Shuford, Albert and Massengill (1966) all argue that differential choice of distractors allows an examinee to exhibit partial information and that this should produce greater item and test variance but should reduce error variance.

Another reason for using a decision-theoretic approach is that by using the concept of a utility function it is possible to specify the types of situations where a student can "rationally" be a risk taker or where he should be.

"honest" in reporting his true beliefs (Roby, 1965; Rippey, 1971; Murphy and Winkler, 1970). The problems associated with guessing and risk taking are not peculiar to probabilistic testing, and if it is argued that in a given testing situation these are important considerations then probabilistic testing yields a conceptual and mathematical format for including them.

To summarize we have listed the following reasons for considering probabilistic testing. First, testing involves making decisions under uncertainty as do many situations faced everyday and as such should be solved by using a "subjective probability decision theoretic" paradigm. Second, using multiple-choice testing situations may be a good way of teaching the subjective probability decision theoretic paradigm. Third, probability testing procedures should lead to more reliable and possibly more valid tests. Fourth, probability testing in conjunction with specific utility functions yields a way of incorporating and handling "risk" and "guessing" behavior in testing situations.

Theoretical and Empirical Findings

This section will attempt to elaborate on the summarization and critique provided by Echternacht (1972). To avoid duplication it is again assumed that the reader is familiar with his discussion (pp. 223-233). Some references not cited by Echternacht are also considered and an annotated bibliography is included as Appendix A.

Echternacht (1972, p. 224) lists six preliminary assumptions underlying probability testing, while Lord and Novick (1968, p. 319) list essentially the same assumptions but distinguish only three assumptions. Since most of the results noted in this section refer to these general assumptions it is worthwhile to quote Lord and Novick's assumptions.

1. The scoring method, as well as the permitted modes of responding, must be known to the subjects. Furthermore subjects must not only know the method but learn to understand fully its implications with particular reference to behavior in the face of uncertainty. Finally they must be able to make the necessary computations to determine an optimal strategy for each item.
2. Examinees must be keenly interested in obtaining a high total score, precisely in the sense of maximizing their total expected score.
3. They must be able to assign numerical values to their subjective probabilities accurately and reliably.

Since in probability testing the examinee is required to specify through his subjective probabilities his degree of belief concerning the various options, the solutions to problems associated with the quantification of these beliefs and their evaluation is central to implementing probability testing.

Van Naerssen (1961) in discussing the measurement of subjective probability was one of the first to note that if the candidates are not informed about the scoring method then the score will depend on the "accidentally chosen strategy" of the candidate. He argues that by telling how many points they can get with each probability rating and explaining that the aim is to get as many points as possible, "a stronger anchoring of the rating categories will be obtained and also a more impartial experiment in which the selectors (are able to) know how they stand" (p. 161). Van Naerssen derives two of the basic scoring rules (logarithmic and quadratic) used in probability testing. Toda (1963) also experimented with these two scoring schemes. Van Naerssen also points out that in deducing these rules it is assumed that the utility of the score is a linear function of the score itself and the effects of non-linearity still need to be determined. Roby (1965) notes that this difficulty is encountered because the person's expressions of his internal belief are influenced by the person's interpretation of where the "payoff lies." Roby shows a possible solution lies in rewarding the person in "direct proportion to the validity of his belief" and that if this is done then the maximum expected value for a person's score occurs when the person "bets" or responds with his true beliefs.

Roby developed a scoring rule for rewarding people which is called the "spherical" scoring system.

Shuford, Albert and Massengill (1966) show that the quadratic, spherical, and logarithmic scoring rules possess the property that an examinee can maximize his expected score on a test (assuming a linear utility functions) if, and only if, he honestly reflected his personal probabilities, that is, when the examinee's expressed probabilities corresponded to his true probabilities. They also point out that with the quadratic and spherical scoring rules the score for any item is determined by the probability assigned to the correct answer and the way in which the student's uncertainty is distributed over the other options. For instance, if (a) is the correct answer to a three option test item then the two responses (.4,.4,.2) and (.4,.3,.3) would yield different scores. However, the scoring rules are "symmetric" in the sense that (.4,.2,.4) would yield the same item score as (.4,.4,.2). On the other hand the logarithmic scoring rule is a function only of the probability assigned to the correct answer. They conclude their arguments for using probabilistic testing by saying:

> In considering the substitution of admissible probability measurement procedures for the choice procedures in current use, it is important to realize that no information will be lost through the substitution since a student's choices can be reconstructed from knowledge of his probabilities and his utility structure with respect to the testing situation. However, the development of appropriate psychometrics and test theory would greatly facilitate the exploitation of the additional information made available through the use of admissible probability measurement procedures. (p. 144)

Winkler (1967a, 1967b) also discusses some problems associated with the quantification of judgment. In his 1967b paper Winkler first notes the distinction between a "good" assessor with respect to the personalistic theory of probability and a "good" assessor who is knowledgeable in the area under consider-

ation. The first context deals with expertise in the general area of probability assessment, while the second context deals with expertise in some area of application. In using probability testing in educational measurement we are trying to reward the most knowledgeable in the second context assuming that the examinee has learned to be a "good" assessor in the first context. These two contexts are identified respectively as the "normative" and "substantive" by Winkler and Murphy (1968). In the normative sense a good probability assessor is one who obeys certain postulates of coherence (consistency) and who expresses probability assessments which correspond to his subjective beliefs or judgments. The actual quantification can be accomplished through using interrogation and bets or through using scoring rules or "penalty functions" which oblige the person, under linear utility, to express his true probabilities. It is the latter that are used in probability testing. Winkler (1967b) doesn't argue that everyone is necessarily a "good" assessor in the normative sense but he does argue that people can be trained to be "good" assessors, he expects people to learn from experience. Training and experience should increase a person's understanding of the methods and fewer inconsistencies should be observed. Training and experience should also lead to a more reliable specification of subjective beliefs into probabilities. That is, naive assessors tend to respond in certain idiosyncratic manners, i.e. they use such numbers as 0, .25, .50, .75 and 1.0 too often or in testing they weight only one or two options.

By comparing a person's assessments and the actual values observed Winkler (1967b) argues that the person can use this information to learn to be a "better" assessor in the second context as well. Such information would be useful to evaluate a person's "bias," i.e. a tendency to consistently underestimate or to consistently overestimate with respect to certain probabilities and situations. Shuford and Massengill (1970) present a way of evaluating such bias for people using their SCoRole.

The assumption of a linear utility function and risk-taking and risk-avoiding are also raised by Winkler (1967b). He noted the problems associated with a non-linear utility but did not present a solution in this paper. He does point out that if ~~Without respect to~~ risk-taking or risk-avoiding ~~exists~~ persists over time, then the person is not following the postulates of coherence or he is operating under some other utility function.

Winkler (1969) points out since probability assessments must be made before the actual outcome is known, then no matter which scoring rule is used, the assessor should maximize his expected score or expected utility. Any of Shuford's et al. "proper" scoring rules can be used in this regard to evaluate assessors in the normative sense. However, the evaluation of assessors in the substantive sense occurs after the outcome is observed. Winkler proves the logarithmic scoring rule is the only one that is compatable with both types of assessments. He also showed that it is possible to relax the assumption of a linear utility function provided you know the form of the non-linear utility function. That is, corresponding to any utility function U and scoring rule S which is "proper" under a linear utility function, it is possible to find a scoring rule which is also "proper" under U. This point is extended further in Winkler and Murphy (1970) and Murphy and Winkler (1970). Also important in the later article is an introduction to sensitivity analysis of scoring rules. That is, how sensitive, in the sense of the scores assigned, are the scoring rules to deviations from optimum assessment of probabilities. The more sensitive the scoring rule the more it "punishes" an assessor as he deviates from reporting his true probabilities. For three values of p (the true probability) they show that in general the logarithmic rule is less sensitive than the quadratic, which in turn is less sensitive than the spherical, although for small deviations all three rules are fairly insensitive. Although they don't mention it, this may be a plus in favor of these scoring rules when used in probability testing. One

objection sometimes given to probability testing is that unless the examinee is an expert in probability assessment you may introduce more error variance through its use than you eliminate. What sensitive analysis might show is that one does not have to assume expertise in probability assessment before using probability testing for "substantive" evaluation.

Staël von Holstein (1970a) notes that the practical uses of scoring rules as feedback devices have been restricted to the areas of meteorology and educational testing. Probability assessment experiments have also been performed in the areas of football (de Finetti, 1962; Winkler, 1967c), stock market prices (Staël von Holstein, 1969) and weather forecasts (Staël von Holstein, 1970b). These experiments all used a quadratic scoring rule and show that it is feasible to obtain probability assessments for non-dichotomous situations. It was not clear from all the experiments whether the subjects in fact became better assessors in the normative sense during the course of the experiments. This was also found in a testing situation by Hansen (1971). In addition Hansen found statistically significant correlations between a measure of degree of certainty in the examinee's responses and the scores on the F scale and Kogan and Wallach risk taking measures. It should be noted that although the correlations were significant they were also relatively low (-.211 to -.411) with most of them below -.250. Hansen used the spherical scoring rule and obtained split-half test reliabilities of .781 and .766 for his two tests.

Phillips (1970) argues that probability judgments can be affected to varying degrees by memory and cognitive processes, prior experience and information, social and cultural norms, personality, and cognitive style. He concludes that to the extent we agree on these variables they should be the focus of future research "since effective training can be designed only when we know how these factors influence the naive person's judgments" (p. 254).

Some other empirical studies done in educational testing are Michael (1968), Rippey (1968, 1970, 1971) and Hambleton, Roberts and Traub (1970). Michael (1968) used the scoring rule $S = r_h$ where $r_h$ is the probability assigned to the correct answer. The probabilities expressed were also restricted to simple tenths. Although she found higher reliabilities and lower standard errors it must be noted that this scoring rule is not "admissible" in that it requires the person not to express his true probabilities when trying to maximize his expected score under linear utility (see Winkler, 1967b, p. 1111).

Hambleton, Roberts and Traub (1970) compared probability testing using a logarithmic scoring rule [possible probabilities were 0, .05, .10, ..., .95, 1.00] with conventional testing and differential weighting. They found probability testing yielded the highest validity (correlation of midterm with final) of the methods (.720) and the lowest split-half reliability (.655). For the conventional test the validity and reliability were .621 and .710 respectively. Two other points of interest in this study are the introduction of an answer graph for reporting probability and mention of the fact that the difficulty of the test will effect the application of probability testing. For instance, the test they used was "easy" for the students involved. In the group using probability testing 77% of the time they indicated a probability of 1.00. In this situation assessing partial knowledge may not be a great concern.

Rippey (1968) applied the logarithmic and spherical scoring rules to the same set of probability responses on a variety of tests and computed the test reliabilities. In comparing these reliabilities he noted that automatic increases in reliability were not found. However, it must be noted the people involved had no experience with probability testing and from the "stereotypical student responses" observed probably would not have passed even a minimum criterion of a good assessor in the normative sense. Another drawback is that as

Winkler (1967b) and Shuford et al. (1966) point out it is important for the person to know and understand the methods being used. In particular the scoring rules may not yield consistent results when applied to the same expressed probabilities. Shuford et al. (1966), as well as Winkler and Murphy (1968), note that the logarithmic rule is concerned only with the probability assigned to the correct outcome, while the spherical and quadratic are concerned with all of the expressed probabilities. However, even these two rules weight the probabilities in different ways. Winkler and Murphy (1968) give a numerical example in which the logarithmic rule yields a higher score for assessor A than assessor B, but if the spherical or quadratic rule is used for the same probabilities assessor B is given a higher score than assessor A. This fact could, indeed, affect the reliability and validity of a test depending upon the scoring rule used. They temper this finding somewhat by noting that they have "evidence that rankings based upon average scores will be reasonably consistent" (p. 756).

Rippey (1970, 1971) reports on another study he completed on the reliability of five different scoring rules. The 1970 reference is a journal article while the 1971 reference is the final report for his USOE grant. The experimental setup was essentially the same as in the 1968 study, in that, it involves naive subjects and applies five scoring rules to the same expressed probabilities. The fact that people might and probably should respond differently under different scoring rules was not considered. The probabilities that the subjects were allowed to use was limited to simple ninth, i.e. 0, 1/9, 2/9, ..., 8/9, 1.0. In his 1970 reference he recommends using the scoring rule $S = r_h$ (see Michael (1968)) since it yields the consistently highest reliabilities although the "Euclidean" rule produced "comparably high reliability." In his (1971) reference Rippey tempers the recommendation for using $S = r_h$ by noting the objection raised above with respect to Michael's (1968) article, and by the fact that his sub-

subjects were naive. If students do learn or are told the optimum strategy for this scoring rule then its use subverts the whole procedure.

The above mentioned literature indicates that a considerable amount of theoretical work has been done. The empirical studies, at least, indicate the feasibility of trying to implement probability testing. Also some of the studies suggest areas in need of more research and it is possible to extrapolate other problem areas from the literature.

## Areas for Further Research

This section of the paper will attempt to list some of the areas for further research that have been identified by the author and others.

Much of the literature reviewed above stresses the importance of training and experience with probability testing in the "normative" sense before it can be used in the "substantive" sense. Some of the research reports mention attempts to familiarize students with the scoring rules, through hypothetical examples, etc.. (Hamilton, et al., 1970; Hanson, 1971). However, one could classify those attempts as orientation rather than deliberate training, in the rigorous sense, with a test for mastery, retention, etc. Phillips (1970) mentions some variables that should be examined in trying to develop training programs in probability assessment and probability testing. In a related context Novick (ACT Technical Bulletin No. 3, no date) has suggested the use of an interactive computer as a strategy for the training of naive people in the area of Bayesian statistical analysis. Rippey (1971) suggests the use of a computer to supply the necessary feedback when using probability testing.

One of Phillips (1970) variables was "personality" and it is also one of the psychological variables needing further study mentioned by Winkler and Murphy (1968, 1970), and de Finetti (1970). Literature concerning personality

characteristics associated subjective probability, risk taking, and decision making are reviewed by Brichacek (1970), Slovic (1964), and Kogan and Wallach (1967).

Winkler and Murphy's (1968) ideas of partitioning assessors into "goodness" categories needs to be extended. One question of interest would be how does "goodness" in the normative sense affect reliability and validity of tests. Closely associated with this is the need for further work in sensitive analysis (see Wagner, 1969) to see how much "expertise" in probability assessment is really needed. They suggest that the sensitivity question may also be related to psychological factors. Much of the experimental work in probability testing has restricted the examinee to limited probability points such as twentieths, tenths, or ninths. Are these too restrictive for probability testing to be effective?

Certainly work needs to be done in developing the "appropriate psychometrics and test theory" to make use of the additional information supplied by probability testing (Shuford, et al., 1966). Since the various scoring rules use the expressed probabilities in different ways, in what testing situations should different scoring functions be used? Also should different procedures be developed for evaluating item discrimination and difficulty. de Finetti (1970) suggests looking at the distribution of probabilities given to the same events by different individuals or groups of individuals. He also suggests that individual scores be compared with the score of a fictitious person "who adopts as his subjective probabilities for each event the average probability given to this event by a group or subgroup. It often happens that this fictitious player is near the top of the performance range" (p. 142).

The implications of non-linear utility functions need more theoretical as well as experimental work. Sensitivity analysis is also applicable here. How

much does the utility function have to deviate from linearity before the expressed probabilities should be shifted from their true values? If we are forcing students into situations necessitating non-linear utility, then should we even be using objective tests no matter how they are administered?

Replications of previous experimental studies with improved procedures should be carried out. As Hambleton et al. (1970) says "Hopefully, other investigators will be stimulated by the inadequacies of the present results to apply the methodology outlined here to investigate what is an important problem in the area of testing" (p. 81).

References

Alberoni, F. Contributions to the study of subjective probability: Prediction. II. *Journal of General Psychology*, 1962, 66, 265-285.

Bakan, D. *On Method: Toward a Reconstruction of Psychological Investigation.* San Francisco: Jossey-Bass, Inc., 1967.

Brichacek, V. Use of subjective probability in decision making. *Acta Psychologica*, 1970, 34, 241-253.

Brunswik, E. Organismic achievement and environmental probability. *Psychological Review*, 1943, 50, 255-272.

Coombs, C.H. On the use of objective examinations. *Educational and Psychological Measurement*, 1953, 8, 308-310.

Coombs, C.H., Milholland, J.E. and Womer, F.B. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, 16, 13-37.

de Finetti, B. Does it make sense to speak of 'good probability appraisers'? In: I.J. Good (ed.), *The Scientist Speculates: An Anthology of Partly-Baked Ideas.* London: Heinemann, 1962, 257-364.

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 1965, 18, 87-123.

de Finetti, B. Logical foundations and measurement of subjective probability. *Acta Psychologica*, 1970, 34, 129-145.

Ebel, R.L. Confidence weighting and test reliability. *Journal of Educational Measurement*, 1965a, 11, 49-57.

Ebel, R.L. *Measuring Educational Achievement.* New Jersey: Prentice-Hall, 1965b.

Echternacht, G.J. The use of confidence testing in objective tests. *Review of Educational Research*, 1972, 42, 217-236.

Good, I.J. *Probability and the Weighing of Evidence.* New York: Hafner, 1950.

Good, I.J. *The Estimation of Probabilities -- An Essay on Modern Bayesian Methods.* Cambridge: MIT Press, 1965.

Hambleton, R.K., Roberts, D.M. and Traub, R.E. A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, 1970, 7, 75-82.

Hansen, R. The influence of variables other than knowledge on probabilistic tests. *Journal of Educational Measurement*, 1971, 8, 9-14.

Hevner, K.A. A method of correcting for guessing in true-false tests and empirical evidence in support of it. *Journal of Social Psychology*, 1932, 3, 359-362.

Hilgard, E.R. The role of learning in perception. IN: R. Blake and G. Ramsey (eds.). *Perception, An Approach to Personality*. New York: Ronald Press, 1951, 95-120.

Jeffreys, H. *Theory of Probability* (3rd ed.). Oxford: Clarendon Press, 1961.

Kogan, N. and Wallach, [illegible] Risk taking as a function of the situation, the person and the group. In: *New Directions in Psychology, III*. New York: Holt, Rinehart and Winston, 1967, 111-278.

Lord, F. and Novick, M. *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley, 1968.

Michael, J.J. The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*, 1968, 5, 307-314.

Murphy, A.H. and Winkler, R.L. Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 1970, 34, 273-286.

Novick, M.R. Bayesian computer-assisted data analysis. *ACT Technical Bulletin*, No. 3, no date.

Phillips, L.D. The 'true probability' problem. *Acta Psychologica*, 1970, 34, 254-264.

Rippey, R.M. Probabilistic testing. *Journal of Educational Measurement*, 1968, 5, 211-215.

Rippey, R.M. A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, 1970, 7, 165-170.

Rippey, R.M. *Scoring and Analyzing Confidence Tests*. Final report of project no. 7-0578, U.S. Department of Health, Education and Welfare, 1971.

Roby, T.B. Belief states and the uses of evidence. *Behavioral Science*, 1965, 10, 255-270.

Shuford, E.H., Albert, A. and Massengill, H.E. Admissible probability measurement procedures. *Psychometrika*, 1966, 31, 125-145.

Shuford, E.H. and Massengill, H.E. *Scalkule Response Aid Instructions*. Lexington: Shuford Massengill Corporation, 1970.

Slovic, P. Assessment of risk-taking behavior. *Psychological Bulletin*, 1964, 61, 220-233.

Soderquist, H.O. A new method of weighting scores in a true-false test. *Journal of Educational Research*, 1936, 30, 290-292.

Staël von Holstein, C. The assessment of discrete subjective probability distributions -- an experimental study. University of Stockholm, Institute of Mathematical Statistics, *Research Report No. 41*, 1969.

Staël von Holstein, C. Measurement of subjective probability. *Acta Psychologica*, 1970a, 34, 146-159.

Staël von Holstein, C. An experiment in probabilistic weather forecasting. Unpublished manuscript, University of Stockholm, Institute of Mathematical Statistics, 1970b.

Toda, M. *Measurement of Subjective Probability Distributions*. ESD-TDS-63-407. Bedford: Decision Science Laboratory, Hanscom Field, 1963.

van Naerssen, R.F. A scale for the measurement of subjective probability. *Acta Psychologica*, 1962, 20, 159-166.

Wagner, H.M. *Principles of Operations Research*. New Jersey: Prentice-Hall, 1969.

Winkler, R.L. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 1967a, 62, 776-800.

Winkler, R.L. The quantification of judgment: some methodological suggestions. *Journal of the American Statistical Association*, 1967b, 62, 1105-1120.

Winkler, R.L. The quantification of judgment: some experimental results. *Proceedings of the American Statistical Association*, 1967c, 386-395.

Winkler, R.L. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 1969, 64, 1073-1078.

Winkler, R.L. and Murphy, A.H. 'Good' probability assessors. *Journal of Applied Meteorology*, 1968, 7, 751-758.

Winkler, R.L. and Murphy, A.H. Nonlinear utility and the probability score. *Journal of Applied Meteorology*, 1970, 9, 143-148.

Appendix A

# Introduction to Confidence Testing: An Annotated Bibliography

Janice Richman, Charles Stegman and Nancy Sorg
Department of Educational Research
University of Pittsburgh, Pittsburgh, Pennsylvania 15213

The following annotated bibliography has been included to introduce potential researchers to the general area of confidence testing. This is only part of a more comprehensive bibliography that is currently being compiled. Copies may be obtained by writing to the authors.

Alberoni, F. Contribution to the study of subjective probability. I. *Journal of General Psychology*, 1962, 66, 241-264.

This is an attempt to determine the psychological meaning of probability. The concepts investigated include the idea of probability and independence. Subjective probability differs from mathematical probability when cause, rather than chance, is suspected to be operating. This may be posited when an order or pattern of some kind emerges in the course of a sample. Another difference is that subjects interpret the probability of a sequence as the probability of that outcome. The subjects are not always coherent.

Alberoni, F. Contribution to the study of subjective probability: Prediction. II. *Journal of General Psychology*, 1962, 66, 265-285.

The psychological processes governing probabilistic prediction are studied. When subjects were asked to supply the next outcome of a sequence of red and blue beads, with an equal number of each color, they used one of three strategies: randomly generating the next outcome with an equal probability of selecting either color, respecting the cyclic nature of the sequence or formally improving the sequence. The latter improvement assumes that the colors in the sequence will alternate in an irregular way. A fourth factor was added when an unequal proportion of the two colors was presented: the quantitative improvement of the sequence. This strategy implies the outcome which best helps the colors in the sequence reflect the proportion in the universe.

Atkinson, J.S., Bastian, J.R., Earl, R.W. and Litwin, G.H. The achievement motive, goal setting, and probability preferences. *Journal of Abnormal and Social Psychology*, 1960, 60, 27-36.

Need for achievement was related to preferences for certain probabilities in a risk-taking model. Those high in need for achievement preferred more intermediate subjective probabilities than those low

In need for achievement, who preferred to set themselves goals with very high (easy shots) or very low (difficult shots) probabilities in an effort to avoid failure. The subjective probabilities were measured in two situations: a shuffleboard game, in which subjects could choose their distance (here, subjective probability was measured geographically!) and in an imaginary betting situation. The preferences did not hold in all the betting situations but only in those with a small monetary reward (30¢).

Boldt, R.F. A simple confidence testing format. ETS Research Bulletin No. 71-42, 1971. ERIC No. ED 056 098.

ERIC Summary: "This paper presents the development of scoring functions for use in conjunction with standard multiple-choice items. In addition to the usual indication of the correct alternative, the examinee is to indicate his personal probability of the correctness of his response. Both linear and quadratic polynomial scoring functions are examined for suitability, and a unique scoring function is found such that a score of zero is assigned when complete uncertainty is indicated and such that the examinee can expect to do best if he reports his personal probability accurately. A table of simple integer approximations to the scoring function is supplied."

Boldt, R.F. An approximately reproducing scoring scheme that aligns random response and omission. ETS Research Bulletin No. 71-43, 1971. ERIC No. ED 057 074.

ERIC Summary: "One formulation of confidence scoring requires the examinee to indicate as a number his personal probability of the correctness of each alternative in a multiple-choice test. For this formulation, a linear transformation of the logarithm of the correct response is maximized if the examinee reports accurately his personal probability. To equate omits scores with choice scores, the transformation can be chosen so that the score is zero if the examinee indicates complete uncertainty. If this is done, the scoring function depends on the number of alternatives. One could also align uncertainty and response omission by granting credit for omitting items, though it is felt this might be hard to explain to examinees."

Cameron, B. and Myers, J.L. Some personality correlates of risk-taking. *Journal of General Psychology*, 1966, 74, 51-60.

The relationships between betting preferences and need states as well as other personality variables are investigated. The betting situation follows the paradigm originated by Ward Edwards, and the Edwards Personal Preference Schedule is the instrument used to measure the personality variables. Betting preferences were measured in both imaginary and actual risk-taking situations, in that order. As in several of Edwards' experiments, probability preferences are confounded with payoff preferences. Subjects high in exhibition, aggression, or dominance tended to prefer bets with high payoff and low probability of winning, while subjects high in autonomy or endurance tended to be more conservative. It is not clear that these five needs on the EPPS are in any way similar to need for achievement as measured by Atkinson et al. (1960).

Coombs, C.H. On the use of objective examinations. *Educational and Psychological Measurement*, 1953, 8, 308-310.

A procedure for administering and scoring objective tests so as to provide a scale from complete misinformation through several degrees of partial information is proposed (Coombs type directions). Individuals should be instructed to cross out all the alternatives they consider to be wrong but not to guess among the remaining options. The weights used in the scoring procedure are as follows: one point is added for each wrong alternative crossed out, k-1 points are subtracted if the right alternative is crossed out (k is the number of options). Advantages of this scoring method are suggested.

Coombs, C.H., Milholland, J.E. and Womer, F.B. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, 16, 13-37.

This study compared conventional test scoring with the scoring procedure outlined by Coombs (1953) in terms of reliabilities, validities and coefficients of discrimination. Positive scores for each item represent some degree of partial information, while negative scores represent some degree of misinformation. Results indicate that examinees with less than complete information on a given subject may have considerable partial information and that this may be used as a valid basis for discriminating among them. The reliabilities were higher for tests administered and scored by the experimental method. This reliability was even further increased for more difficult tests. Both types of scoring appear to be equally valid. What constitutes a good discriminating item is the same for both methods.

Coombs, C.H. and Pruitt, D.G. Components of risk in decision making: Probability and variance preferences. *Journal of Experimental Psychology*, 1960, 60, 265-277.

An alternative to Ward Edwards' theory of maximization of subjectively expected utility is proposed. This model involves variance preferences, as well as probability, skewness and expectation preferences. An experimental betting situation supports the hypothesis that variance preferences exist and can be generated by folding a joint scale. However, for each set of variance preferences, a nonlinear utility function of money can be found which explains the ordering equally well. Skewness preferences were also found to exist. One conclusion was the subjects are inconsistent in their preferences.

Dale, H.C.A. A study of subjective probability. *British Journal of Statistical Psychology*, 1960, 13, 19-29.

Adult subjects' predictions of how a small number of items would be selected by chance from a long list were compared to the objective probabilities. The subjects appeared to avoid unlikely configurations but did not consider all the aspects of the selection process that the authors had determined were important a posteriori. Three aspects of configurations were chosen for consideration: range, bunching and symmetry. None of the models proposed seemed to adequately describe the subjects' behavior.

Davis, F.B. Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, 1959, 19, 291-298.

If the options of a multiple-choice item are to be weighted according to their degree of correctness, the appropriate weights remain to be determined. By administering the items to a large representative sample, a scoring weight for each item option can be found that is linearly related to the average score on the criterion variable of those in the try-out sample who selected that choice. Since direct computation of the average criterion score for the group selecting each option is very time consuming, a method of estimating the criterion score means is given in tabular form, requiring only the percent of those in the upper 27% and in the lower 27%, respectively, who selected the given option. The estimated means were found to produce moderately reliable weights and very close to the weights calculated by the actual criterion-score means.

Davis, F.B. and Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, 19, 159-170.

It was found that scoring an arithmetic reasoning test by weighting the options according to their degree of correctness was more reliable than conventional scoring. The validity of the test was unaffected. Weights were assigned in three ways: a priori weights were determined independently by two mathematicians, empirical weights were obtained by using a function of the average criterion scores of those selecting each choice for a previous group of examinees who took the test scored by a priori weights, and modified empirical weights were approximated from the scores of the upper and lower 27% of the previous sample. a priori weights seem to be a necessary feature in determining the subsequent empirical weights. Otherwise both kinds of empirical weights may actually be based on differential appeal of the wrong options rather than degree of correctness and thus may not be assessing partial knowledge.

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 1965, 18, 87-123.

In the absence of complete information a person should be encouraged to attach a probability to each alternative. This probability should correspond to the individual's degree of belief as to the correctness of that alternative. Other answering techniques are discussed, including Coombs' type of directions. All techniques are interpreted geometrically. Subjective probability leads to a scoring system that makes sense, unlike the rank ordering or crossing out of a number of wrong alternatives. Training in the use of a suitably selected technique is recommended. A strong case is made for assessing and utilizing partial knowledge in scoring multiple-choice questions.

de Finetti, B. Logical foundations and measurement of subjective probability. *Acta Psychologica*, 1970, 34, 129-145.

Subjective probability is considered the only meaningful interpretation of probability. Terms involving properties associated with objective probabilities, such as event and stochastic independence, should be avoided. Probability is degree of belief and must be operationally defined by some device such as offering a suitable set of bets, fixing a penalty, or introducing an opponent. Probabilities must be consistent to be admissible; however, logical or empirical consideration may suggest further restrictions. Scoring rules are briefly discussed. Ten psychological criteria for evaluating assessors are outlined. Recourse to concepts of "objective probability" is examined and rejected.

Dressel, P.L. and Schmid, J. Some modifications of multiple choice items. *Educational and Psychological Measurement*, 1953, 13, 574-595.

Five scoring methods are compared in terms of their reliabilities: free choice, in which any number of options can be selected only one of which is correct; degree of certainty, in which the student marks how certain he/she was about the option selected on a scale of 1 to 4; multiple answer in which any number of options can be selected and more than one option may be correct; two-answer, in which two options are correct; and a conventional test. The highest reliability was found for the multiple-answer test. The two-answer and degree of certainty tests had slightly higher reliabilities than the conventional test.

Ebel, R.L. Confidence weighting and test reliability. *Journal of Educational Measurement*, 1965, 2, 49-57.

A system of confidence-weighted response and scoring was developed for true-false test items. A justification for the use of the true-false format in high quality tests of educational achievement is given. Previous data had shown that tests weighted by confidence had significantly higher reliabilities than conventional tests. Recent data, however, showed a negligible increase in reliability for the weighted scoring, not enough to justify the more complicated technique. Simulating a set of responses and scoring by weighted and conventional techniques suggests that confidence weighting should only be applied to those items with a higher than chance probability of a correct response (the criterion used in the simulation was two-thirds).

Ebel, R.L. Review of "Valid confidence testing -- demonstration kit." *Journal of Educational Measurement*, 1968, 5, 353-354.

This is a review of Shuford-Massengill's materials for Valid Confidence Testing, which include: SCoRule response aid, answer sheets, a scoring table, and a class analysis form. The process seems complex, and the costs seem high. Indirect evidence as to the indicated degrees of confidence being related to the proportion of correct answers is given. Valid confidence scores correlate substantially, but not perfectly, with conventional scores. There is only incomplete support for Shuford and Massengill's claims of increased reliability and validity.

Echternacht, G. et al. User's handbook for confidence testing as a diagnostic aid in technical training. ETS Report No. PR-71-12, 1971, ERIC No. ED 055 119.

ERIC Summary: "This handbook presents instructions for implementing a confidence testing program in technical training situations, identification of possible areas of application, techniques for evaluating confidence information, advantages and disadvantages of confidence testing, time considerations, and problem areas. Complete instructions for "Pick-One" and "Distribute 100 Points" confidence testing methods are given for testing supervisors and examinees for both hand and computer scoring."

Echternacht, G. The use of confidence testing in objective tests. ETS Research Bulletin No. 71-41, 1971. ERIC No. ED 058 307.

ERIC Summary: "Confidence testing has been used in varying forms over the past 40 years as a method for increasing the amount of information available from objective test items. This paper traces the development of the procedure from Hevner's beginning method up to the various methods in use today and describes both the testing procedures and scoring methods used. The term confidence testing is applied to both probabilistic testing and confidence weighting procedures. Various procedures are presented and their relationship with personality factors discussed."

Echternacht, G.J. The use of confidence testing in objective tests. *Review of Educational Research*, 1972, 42, 217-236.

Various forms of confidence testing are described and evaluated. In spite of Jacob's distinction (1971) between confidence weighting and probabilistic testing, they are here subsumed under one rubric, that of confidence testing. The sole use of the criterion of increasing reliability in evaluating confidence testing is criticized.

Garvin, A.D. Confidence weighting. Paper presented at the annual meeting of the American Educational Research Association, 1972. ERIC No. ED 062 401.

ERIC Summary: "Various aspects of Confidence Weighting are examined. Variant of Confidence Weighting, its effect on test reliability, and the validity of Confidence Weighting are discussed."

Hambleton, R.K., Roberts, D.M. and Traub, R.E. A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, 1970, 7, 75-82.

Three groups were compared on the basis of different instructions and scoring methods: conventional method, differential weighting of distractors according to the degree of correctness (determined by 22 experts), and confidence testing using an answer graph and a reproducing logarithmic scoring function. Confidence testing was most valid and least reliable. Validity was determined by correlating the scores with midterm scores. Reliability was estimated from corrected split-half correlations, a method that has been considered by some to be inappropriate for confidence testing. Two sets of differential weights were developed from the experts' ranking, one considerably more complex than the other. The more complex weights were more valid and less reliable than the simpler weights. The simpler weights were as reliable as conventional scoring and more valid. A more difficult test might have proved more informative.

Hansen, R. The influence of variables other than knowledge on probabilistic tests. *Journal of Educational Measurement*, 1971, 8, 9-14.

Individuals who take examinations using a probabilistic scoring system display a relatively stable tendency which cannot be accounted for on the basis of their stability of knowledge. The tendency of an individual to show certainty was determined from a function of the probabilities assigned to the options. This measure is highest where certain options are assigned probability of 1 and lowest when the probabilities are equally distributed among the options. The test score was computed using the spherical scoring function. The correlation between the measures of certainty for two successive exams was .702. The correlations between test score and the measure of certainty were low. On the other hand, this tendency correlated positively with Kogan and Wallach's measure of risk-taking, the Choice Dilemma Questionnaire and negatively with the F-scale. Both correlations were moderate (less than .42).

Hopkins, K.D., Hakstian, A.R., and Hopkins, B.R. Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement*, 1973, 33, 135-141.

Confidence weighting studies are summarized in tabular form and are shown to have resulted generally in somewhat higher reliabilities. Three studies using subjective probability are subsumed under the C.W. rubric. The gain in reliability is hypothesized to be a result of a gambling response style, or irrelevant source, in which case a decrease in validity might occur. A final exam was administered with confidence weights, of the form high, medium and low. An item score could range from -3 to +3. A short answer exam on the same material provided the validity criterion. Conventional scoring resulted in slightly higher validity and lower reliability than confidence weighted scoring. The authors conclude that the added variance in the confidence weighting studies may be irrelevant response style variance since validity was not increased.

Liverant, S. and Scodel, A. Internal and external control as determinants of decision making under conditions of risk. *Psychological Reports*, 1960, 7, 59-67.

Internal versus external control is found to be another personality variable entering into making risky decisions. Internal-external control is a construct which depends on whether an individual categorizes desirable and/or undesirable items as within or beyond his control. The I-E scale used is an extension of work done by Games (1957). A betting situation in which individuals can choose between bets differing in pay-off confounded with probability was set up. It was hypothesized that internally controlled persons would tend to employ a strategy which would attempt to maximize the number of favorable outcomes. Externally controlled people would be disposed to select bets more subjectively, on the basis of "hunches" or the outcome of previous trials. The I's did choose more immediate and fewer low probability bets than the E's. Significantly more I's than E's never selected an extreme high or low probability bet. The amount of money wagered on safe, as opposed to risky, bets was greater for I's.

Marschak, J. Actual versus consistent decision behavior. *Behavioral Science*, 1964, 9, 102-110.

General hypotheses of decision behavior are suggested to explain how people make decisions when the problem is too complex for them to apply the utility principle. These hypotheses include "rational" or "consistent" behavior, learning theory, stochastic decision theory, applying Gestalt theory, and the effect of training. Experiments are proposed to determine whether subjects are applying the principles of expected utility, namely consistency, admissibility, independence.

Michael, J.J. The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*, 1968, 5, 307-314.

The reliabilities and standard errors of measurement were compared for the methods of scoring the same test: conventional scoring, the number right corrected for guessing, and confidence weighting. In the confidence weighting method ten points were to be distributed among the four alternatives. That method had the highest reliability and lowest standard error of measurement of the three. The reliabilities broken down by sex and IQ were only slightly different under confidence weighting.

Murphy, A.H. and Epstein, E.S. Verification of probabilistic predictions: a brief review. *Journal of Applied Meteorology*, 1967, 6, 748-755.

The evaluation process is defined as one consisting of several ordered steps. The first step is to identify the purposes of evaluation, which in this article lead to distinguishing between two forms of evaluation: operational evaluation, which is concerned with the value to the user of probabilistic predictions, and empirical evaluation or verification, which is concerned with how closely the predictions correspond to actual observations. Desirable properties for empirical evaluation are enumerated as perfection and unbiasedness and compared with terminology adopted by other authors. Seven measures or scores of the properties are considered, including probability scores, information ratios, and distance measures. Two prediction systems are compared on the basis of different measures.

Murphy, A.H. and Winkler, A.L. Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 1970, 34, 273-286.

Scoring rules are discussed in the contexts of probability assessment, in which the expected scores are of interest, and evaluation, in which the "goodness" of the probabilities should be measures. Scoring rules to be used in assessment should encourage the assessor to be honest in reporting probabilities. If the assessor has a linear utility function, scoring rules should be sensitive to deviations of expected scores from the probability judgments. Four scoring functions (logarithmic, quadratic, spherical and ranked probability score) were compared in a few cases as to sensitivity. No conclusions as to which was most sensitive could be drawn, although the logarithmic function appeared least sensitive. With a nonlinear utility function which is unknown and cannot

be incorporated into the scoring rule, the assessor's statements may differ from the assessor's actual judgments. Several frameworks for evaluation were described. From the inferential viewpoint validity, or the association between the probability statements and the actual outcomes was most important. Roberts' Bayesian model using likelihood ratios was mentioned, as were decision theoretic frameworks.

Pascale, P. Innovation in item scoring procedures, 1971, ERIC No. ED 056 096.

ERIC Summary: "This brief review explains some alternate scoring procedures to the classical method of summing correct responses. The novel procedures attempt in some way to retrieve and use even the information in the wrong responses."

Ramsay, J.O. A scoring system for multiple choice test items. *British Journal of Mathematical and Statistical Psychology*, 1968, 21, 247-250.

If the purpose of a multiple-choice test is to classify an individual into one of two groups, each alternative or option can be weighted by the differences between the probabilities of selecting that option for two criterion groups. Scores weighted in this fashion maximize the separation between the mean scores of the two criterion groups. The results are extended to more than two criterion groups. Advantages of this scoring system are that partial knowledge is taken into account, computations are minimized, item selection is enhanced, and reliability is expected to be improved. Disadvantages are that the system does not imply that misclassifications are minimized and that it may indeed perpetuate any initial miscalssification.

Rippey, R.M. Probabilistic testing. *Journal of Educational Measurement*, 1968, 5, 211-215.

Four tests were administered and scored probabilistically to determine whether increases in reliability would result. Two scoring functions were used: spherical and logarithmic. An increase in reliability was observed in the first test coupled with a corresponding increase in administration time. Different items would be retained in the probabilistic case on the basis of item analyses. Stereotypical student responses were observed, indicating that students may have trouble in thinking probabilistically with respect to more than two classifications. The probabilistic score correlated lower with an essay test on the same material than did conventional scoring. In general, the results were anomalous.

Rippey, R.M. A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, 1970a, 7, 165-170.

Five probabilistic scoring functions were compared on the basis of their reliabilities. All five functions were applied to the same tests. The functions were: probability assigned to correct answer, logarithmic, spherical, Euclidean, and inferred choice. The simplest function, the probability assigned to the correct choice, proved the most reliable. Inferred choice, which is equivalent to conventional scoring, was least reliable. The conclusions were that the simplest and most intuitive scoring functions were best since they were most likely to correspond to the expectations of the examinees.

Rippey, R.M. Rationale for confidence-scored multiple-choice tests. *Psychological Reports*, 1970b, 27, 91-98.

If subject responses related to incomplete information, uncertain knowledge, or degree of preference are to be sampled, confidence-scoring procedures for conventional items or the use of intrinsic items is recommended. Intrinsic items require a distribution of belief over the options on a multiple-choice test and do not have unique correct responses. A Euclidean scoring function scores intrinsic items on the basis of the distance between the probabilities of the individual and the criterion group mean for each response. Since items which call for uniform distributions of confidence over all responses may not discriminate between the informed and the uninformed, a confidence weight on the assigned distribution of belief is suggested.

Rippey, R.M. Scoring and analyzing confidence tests. Final report of project no. 7-0578, U.S. Department of Health, Education and Welfare, 1971.

The literature leading up to and including probabilistic testing is reviewed. New features include an entropic scoring function and a Euclidean function weight by degree of confidence. Three tests with non-unique correct answers were devised and scored with the weighted and unweighted Euclidean functions. Confidence was extensively correlated with sex, grade, and socioeconomic class.

Roby, T.B. Belief states and the uses of evidence. *Behavioral Sciences*, 1965, 10, 255-270.

A new notation called B-state or belief state is introduced to facilitate updating prior beliefs with current evidence. Advantages of this approach are that (1) quantitative comparison or combination of the beliefs of several individuals or one individual at several time periods is possible and (2) the effects of external evidence can be described as mathematical operations on the existing belief state. With the necessity for absorbing new notation, it is not clear that the B-state operators are superior to Bayes' theorem.

Romberg, T. et al. Three experiments involving probability measurement procedures with mathematics test items. Wisconsin Research and Development Center for Cognitive Learning Report No. Tr-129, 1970. ERIC No. ED 044 315.

ERIC Summary: "This is a report from the Project on Individually Guided Mathematics, Phase 2 Analysis of Mathematics Instruction. The report outlines some of the characteristics of probability measurement procedures for scoring objective tests, discusses hypothesized advantages and disadvantages of the methods, and reports the results of three experiments designed to learn more about the technique and compare it with standard procedures of scoring objective tests. The procedure used required the students to specify a degree of belief probability for each of the given alternatives to a question. The students were given a multiple-choice item and asked to specify what they believed to be the probability of correctness

of each choice. The initial intent of these experiments was to see if a non-standard test-taking and scoring procedure would provide useful, reliable information for such tests. The studies indicated that the problem of getting useful, reliable information on difficult tests has not been solved."

Scodel, A., Ratoosh, P. and Minas, J.S. Some personality correlates of decision-making under conditions of risk. *Behavioral Science*, 1959, 4, 19-28.

Personality variables are incorporated into the utility-maximization model. Risk taking was measured in a gambling situation following Edwards' paradigm, in which probability preferences and payoff preferences were similarly confounded. The college grouped tended to be more conservative than the military group. Intelligence was inversely related to variability in risk-taking, but not related to degree of risk-taking. The group choosing low payoffs had more fear of failure and less need for achievement than the high or intermediate payoff groups.

Shuford, E.H., Albert, A. and Massengill, H.E. Admissible probability measurement procedures. *Psychometrika*, 1966, 31, 125-145.

A probabilistic scoring system for objective tests which allows the student to maximize his/her expected score if and only if he/she honestly reports the degree-of-belief probabilities which should have the reporducing property. Necessary and sufficient conditions for the scoring system to have a reproducing property are stated and proved. A method is given for generating a class of functions, both symmetric and asymmetric, possessing the reproducing property. Scoring systems are chosen which reward intelligent probability assessments: the more probability placed on the correct option, the higher the score. With a minor modification the results can be extended to testing situations in which the student has to generate the answers as well as indicate degree of belief.

Slakter, M.J. Riak taking on objective examinations. *American Educational Research Journal*. 1967, 4, 31-43.

A model of risk-taking on objective examinations under conventional directions is included. Measures of risk-taking used in the past are reviewed, including Swineford's gambling terdency, the number of omitted responses, and Coombs' type directions. A new measure of risk-taking is proposed. Coombs' type directions are given, and a number of nonsense questions are inserted into the test. An index is defined, based on the number of alternatives in the nonsense questions which are crossed out. A correlational study showed the new measures of risk-taking to be reliable. Some evidence for convergent and discriminant validity is offered.

Slakter, M.J. Generality of risk-taking on objective examinations. *Educational and Psychological Measurement*, 1969, 29, 115-128.

The question of whether risk-taking on objective tests is a general phenomenon which applies to various kinds of testing situations is examined. The measure of risk-taking involved imbedding nonsense questions in the test. The generality of the risk-taking factor was supported by the correlations bteween the risk-taking measures for four tests: mathematics, language, aptitude and achievement.

Slovic, P. Convergent validation of risk-taking measures. *Journal of Abnormal and Social Psychology*, 1962, 65, 68-71.

The intercorrelations among several risk-taking measures of different kinds were examined to determine whether they were high enough to provide support for convergent validity. The response set measures included the Dot Estimation test, which reflected speed versus accuracy; Word Meanings, which measured inclusiveness of category width; and Test Risk used a variant of Coombs' type directions and accounted for gambling set. Questionnaires used were the Life Experience Inventory and the Job Preference Inventory. Experimental gambling measures were taken with the Bet Preference and the Self-Crediting test, both of which investigated variance preferences, low intercorrelations (below .35) indicate a lack of convergent validity.

Slovic, P. Assessment of risk-taking behavior. *Psychological Bulletin*, 1964, 61, 220-233.

The literature relevant to the validity of various risk-taking measures is extensively reviewed. The studies are classified into three categories: response set and judgmental measures, questionnaire measures, and probability and variance preference measures. The lack of agreement in convergent validity might be due to the multidimensionality of risk, the subjectivity involved in perceiving risk, or the emotional or autonomic response necessary to arouse risk-taking tendencies. The bibliography is very inclusive.

Staël von Holstein, C.-A.S. Measurement of subjective probability. *Acta Psychologica*, 1970, 34, 146-159.

Scoring rules are discussed in a highly understandable manner. Proper scoring rules and strictly proper scoring rules are defined. Criteria for selecting one scoring rule over another are mentioned. These include Raiffa's principles of relevance, univariance and strong discriminability. Roberts' Bayesian model for comparing probabilistic predictions is shown to invoke these three principles. A scoring rule is developed that is sensitive to distance, or orderings of the possible events. This rule conflicts with Raiffa's principles. Practical uses of scoring rules as feedback devices are presently restricted to the areas of meteorology and educational listing. Assessment techniques not based on scoring rules are briefly reviewed, including Winkler's questionnaire which uses four methods to elicit underlying distributions. Toda's "range betting method" is mentioned.

Stanley, J.C. and Wang, M.D. Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, 1970, 30, 21-35.

The literature encompassing differential weighting of items as well as options is reviewed. Differential weighting of items with the same weights for all examinees seems useless. However, two modifications seem promising. Birnbaum different tally weighted items by the levels

of ability of the examinees, and Cleary developed a procedure for using individual regression weights. Differential weighting of options was originally developed to maximize the relationship of the instrument with outside criteria. Guttman keyed each option against a quantitative criterion using the criterion mean of those who chose that option as the scoring weight. A cursory review of the personal probability weightings of the options is presented, and the approach is recommended with modifications.

Swineford, F. Measurement of a personality trait. *Journal of Educational Psychology*, 1938, 29, 295-300.

The tendency to gamble, a personality trait affecting objective test scores, is measured by incorporating an instruction into the testing situation whereby the student can claim from two to four points credit for each item. The student is penalized by double the amount of credit claimed if the wrong option is chosen. The gambling score is the percentage of errors marked "4" to the total number of error plus one-half of the omissions for a true-false test. The gambling score formula yields a reliable measure of a trait which is independent of achievement on the same test. The test should be difficult for this measure to be reliable.

Swineford, F. Analysis of a personality trait. *Journal of Educational Psychology*, 1941, 32, 438-444.

The tendency to gamble was measured on four tests administered to the same population. One fourth of the 457 students were eliminated from consideration since on at least one test either no extra credits were claimed or no errors were made. In either case no gambling score could be computed. Boys exhibited a significantly higher tendency to gamble than girls, especially on unfamiliar types of tests. More students gambled on unfamiliar material. The gambling scores were in most cases independent of five mental factors and correlated highly with each other.

van Naerssen, R.F. A scale for the measurement of subjective probability. *Acta Psychologica*, 1962, 20, 159-166.

To avoid measuring subjective probability by the more cumbersome method of paired comparisons, the subject or selector has to choose between a number of ordered pairs at the same time. A type of scale is developed with a quadratic solution. Applications are the measurement of subjective probabilities as in assessing level of aspiration or predicting success or failure for a candidate and the increasing of the reliability of two choice tests.

Winkler, R.L. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 1967a, 62, 776-800.

University of Chicago students were questioned using four techniques: Cumulative Distribution Function, Hypothetical Future Samples, Equiva-

lent Prior Sample Information, and Probability Density Function, in order to elicit enough information to write down their prior distributions. Subjects had trouble with CDF's but in general learned to assess prior distributions on their own. A revised questionnaire is presented in the appendix.

Winkler, R.L. The quantification of judgment: some experimental results. Proceedings of the American Statistical Association, March 1967b, pp. 386-395.

The efficacy of scoring rules and bets in keeping assessors of subjective probabilities honest and providing them with feedback is investigated experimentally. The 13 week study involved the weekly assessments of various probabilities and the expected point spread of weekend football games. The subjects were given feedback from two scoring rules, the quadratic evaluating their probabilities and a squared-error loss evaluating their spread. They were then given a chance to make bets on the basis of their probability assignments. The scoring rules and bets seemed to lead the assessors to make careful assessments. A consensus of assessors compared favorably to the performance of the individuals comprising the consensus.

Winkler, R.L. The quantification of judgment: some methodological suggestions. *Journal of the American Statistical Association*, 1967c, 62, 1105-1120.

An ideal assessor of personal probability, who never violates the postulates of coherence, is imagined to be faced with choices of bets. In order to force true responses as to his personal probabilities, a penalty or scoring function must encourage revelation of the probabilities. Four proper scoring rules are described: de Finetti's rule, the "Brier score," the spherical gain, and the logarithmic loss. The implications and practicality of these methods are discussed.

Winkler, R.L. and Murphy, A.H. "Good" probability assessors. *Journal of Applied Meteorology*, 1968, 7, 751-758.

A framework for evaluating meteorologists who assess probabilities must be consistent with the theory of subjective probability. Two standards of "goodness" are described normative, which requires the assessor to obey thepostulates of coherence and make honest assessments, and substantive, concerned with knowledge of the subject and reflected in the degree of association between the predictions and the observations. Three proper scoring rules are discussed quadratic, spherical, and logarithmic. The logarithmic scoring rule only considers the probability of the outcome that occurs, while the other two are concerned with all the probabilities. Proper scoring rules encourage assessors to be honest, permit evaluation of assessors, and help individuals become better assessors. Proper scoring rules may not yield consistent results, since they may not assess the same aspects of the attribute validity. Rankings based on average scores may be reasonably consistent.

Winkler, R.L. and Murphy, A.H. Nonlinear utility and the probability score. *Journal of Applied Meteorology*, 1970, 9, 143-148.

Proper scoring rules assume that the assessor has a linear utility function. If the utility function is actually nonlinear, as in the cases of a risk-taker and a risk-avoider, factors other than the expected score may affect the probability forecasts. The expected utility is found to depend on the variance of the score as well as the expected probability score for the risk-taker. The optimal forecast for an extreme risk-taker would be to assign the event probability one if the assessor's actual subjective probability, $p_i$, were greater than one-half and zero if $p_i$ were less than one-half. A risk-avoider is presumed to prefer a small variance to a large one. An extreme risk-avoider would prefer probabilities close to one-half. If the assessor's utility function can be specified, it should be incorporated into the assessment process by defining a new rule, a composite of the original rule and the utility function. If the utility function cannot be determined, the assessor's statements may differ from the true subjective probability judgments.

Ziller, R.C. A measure of the gambling response-set in objective tests. *Psychometrika*, 1957, 22, 289-292.

A formula for measuring risk-taking or gambling set in objective tests is developed. The index of risk-acceptance depends on the number of alternatives the number of incorrect responses, and the number of omissions. The index is designed for tests in which examinees are informed that a correction for guessing will be applied. A few implications of this measure for test theory and construction are discussed.