EVALUATION IN EDUCATION:
AN INTERNATIONAL REVIEW SERIES

Confidence Marking: Its Use in Testing

The Author

DIEUDONNÉ LECLERCQ. Born in 1944 in Belgium, received his doctorate in education in 1975 at the University of Liège. He is an Associate Professor at that University in the Laboratoire de Pédagogie Expérimentale where he teaches educational technology. He has published several books and many articles – all in french – on programmed learning, item banking, and multiple choice items.

Pergamon Reviews in Educational Evaluation

This series comprises the two review journals *Evaluation in Education* and *Studies in Educational Evaluation* and covers all aspects of educational evaluation, including curriculum evaluation, studies of educational systems and educational organizations, evaluation of teaching/learning strategies and assessment of student performance. These two journals may be subscribed to independently or at a reduced combined subscription rate, and are also available as single issues.

ISBN 0 08 031047 8
ISSN 0191–765X
EIRSD6 6 (2) 161–288 (1982)

LOC 77–81507

Volume 6 Number 2 1982

Evaluation in Education: An International Review Series

CONTENTS

PERGAMON PRESS

OXFORD · NEW YORK · TORONTO · PARIS · FRANKFURT · SYDNEY

Volume 6 Number 2 Pages 161–288 1982

Evaluation in Education: An International Review Series 1982

EVALUATION IN EDUCATION: AN INTERNATIONAL REVIEW SERIES

Editors:

Professor Bruce H. Choppin, *Center for the Study of Evaluation, Graduate School of Education, Moore Hall, UCLA, Los Angeles, California 90024, U.S.A.*

Professor T. Neville Postlethwaite, *Department of Comparative Education, University of Hamburg, Sedanstrasse 19, D2000 Hamburg 13, Germany*

Editorial Advisory Board:

Dr Hiroshi Kida, *Director General, National Institute for Educational Research, 5-22 Shimomeguro, 6-Chome, Megero-Ku, Tokyo, Japan*

Professor A. A. El Koussy, *10 Moh Kamel Mursi St., Dokki, Giza, Egypt*

Professor B. McGaw, *Professor of Education, Murdoch University, Murdoch, Western Australia 6153*

Dr Moegiadi, *Seames, 920 Sukhumvit Road, Bangkok 10110, Thailand*

Professor H. J. Walberg, *College of Education, University of Illinois at Chicago, Box 4348, Chicago, Illinois 60680, U.S.A.*

Evaluation in Education: An International Review Series is published as 1 volume of 3 parts per year. 1984 Subscription rate: $60.00 (including postage and insurance). 1984/1985 rate $114.00.
1984 Combined subscription rate for *Evaluation in Education* and *Studies in Educational Evaluation*: $105.00 (including postage and insurance).

Specially reduced rate to individuals. Any individual whose institution takes out a library subscription may purchase a second or additional subscription for personal use at a much reduced rate of $25.00 per annum.
Single back volumes and parts are available. Prices of single parts are available on request. *Subscription enquiries from customers in North America should be sent to* Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, NY 10523, U.S.A., *and for the remainder of the world to* Pergamon Press Ltd, Headington Hill Hall, Oxford OX3 0BW, U.K.

Microform subscriptions and back issues

Back issues of all previously published volumes are available in the regular editions and on microfilm and microfiche. Current subscriptions are available on microfiche simultaneously with the paper edition and on microfilm on completion of the annual index at the end of the subscription year.

Publishing/Advertising Offices
Pergamon Press Ltd
Headington Hill Hall
Oxford OX3 0BW, England

Pergamon Press Inc
Maxwell House
Fairview Park, Elmsford
New York 10523, U.S.A.

Evaluation in Education, Vol. 6, pp. 161–287, 1983
Printed in Great Britain.

0191-765X/83 $0.00 + .50

CONFIDENCE MARKING: ITS USE IN TESTING

Dieudonné Leclercq

Laboratoire de Pédagogie Expérimentale, University of Liège, Belgium

CONTENTS

CHAPTER 1

MODELS OF MENTAL ACTIVITY AND THE EXPRESSION OF DOUBT

INTRODUCTION

The use of confidence indices in educational settings has a long history (see Henmon, 1911; Hollingworth, 1913; Trow, 1923; Hevner, 1932; Jacobs, 1968; Allgren, 1967).

As early as 1906, Cooke, the Government Astronomer for Western Australia, advocated that each metereological prediction be accompanied by a single number which would "indicate, approximately, the weight or degree of probability which the forecaster himself attaches to that particular prediction." He reported (Cooke, 1906a, 1906b) results from 1,951 predictions. Of those to which he had attached a weight of 5 ("almost certain to be verified"), 98% were correct. For his weight of 4 ("normal probability"), 94% were correct, while for his weight of 3 ("doubtful"), 77% were correct.

This example, reported by Lichtenstein and others (1977), does not emphasize the "scoring" aspect of confidence marking. Ebel's definition (1965a, p. 49) is a good starting point to show the various connections between confidence marking and models of mental processes as well as the accompanying theoretical and mathematical aspects of psychometrics.

According to Ebel, "the term confidence weighting refers to a special mode of responding to objective test items, and a special mode of scoring those responses. In general terms, the examinee is asked to indicate not only what he believes to be the correct answer to a question, but also how certain he is of the correctness of his answer. When his answers are scored, he receives more credit for a correct answer given confidently than for one given diffidently. But the penalty for an incorrect answer given confidently is heavy enough to discourage unwarranted pretense of confidence ".

Some remarks should be made:

a) the expression "objective test" is inappropriate if it is interpreted only in terms of multiple choice questions. Historically, the debate on scoring multiple choice items evoked confidence marking because it was a procedure that could solve the problem. But, theoretically, confidence

marking can be applied equally well to open ended items and restricted choice items.

b) Confidence marking is a way of expressing an individual's subjective probability that his answer will be considered as correct by the teacher (or the corrector, or the mechanical scoring system). The expression *degrees of confidence* is the most frequently used, but various other expressions will be found as well: degrees of assurance, of certainty, of sureness, of conviction (Irwin, 1973). Nevertheless, as will be seen, all these expressions should refer to a probability scale or to the equivalent scale of percentage of chances.

c) Some authors suggest distinguishing between two major kinds of confidence marking:
   - *probabilistic testing*, where the student has to indicate the degree of confidence he gives to each alternative (for a multiple choice item) or each possible answer (for an open ended item).
   - *confidence weighting*, where the student has to give *only one* answer and one confidence degree associated with it.

However, there is no agreement about this distinction or the need to adopt it. Therefore, in the following pages, confidence or probabilistic testing, weighting or marking will be considered as equivalent terms.

The Need for Models

Consciously or not, each testing and scoring procedure is related to an epistemological model, that is, to a theory of knowledge, and what is a relevant measure of it. In multiple choice scoring, the models referred to frequently include decision processes. The choice of a confidence degree (amongst the available ones) is a situation of decision making. *Decision theory* is therefore helpful. Technical problems like validity, reliability, acuity, etc. could not be soundly treated without a serious study of the relations between confidence marking, models of mental activity, decision theory, the planning of intructions and the choice of appropriate tariffs.

Scandura's witty remark (1977) must be interpreted seriously: "Nothing is more practical than a good theory" (provided that a good one exists).

For instance, the various procedures to cope with guessing (the so-called corrections for guessing) can be related to three categories of models. Undoubtedly, the majority of authors who recommended or used such corrections were probably not conscious of these underlying models.

An examination of the evolution of these three models is the most helpful way to show the modification of the use of confidence marking procedures. Let us focus on the classical situation where a student has to answer a multiple choice question by selecting only one answer, it being assumed that there is one (and only one) good or correct answer among the alternatives. What is the student's mental activity?

Choppin (1971) has described three models that have been used to represent this mental activity associated with several procedures of testing and scoring. This chapter will essentially develop and criticize the three models and their implications. In particular, attention should be paid to various *scales of tariffs* (points attributed for success, for failure, and for omission). These scales of tariffs will be referred to as *t scales*.

The best known t scale will be referred to as the St scale (simple t scale) where TC = +1 (tariff for a correct answer), TI = 0 (tariff for an incorrect answer) and TO = 0 (tariff for omission).

THE FIRST MODEL OF MENTAL ACTIVITY IN MULTIPLE CHOICE ANSWERING

The statement of Model 1 (cf. Choppin, 1971)

- When a student knows the correct answer, he chooses the corresponding alternative.
- When he does not know the correct answer, he guesses wildly (randomly) among the alternatives.

In the second statement above, the student's probability (p) of getting the correct answer is 1/k where k *is the number of alternatives*:

$$p(rg) = \frac{1}{k}$$

In the above statement, p(rg) means "p (in case of random guessing)".

Model 1 and the classical correction for guessing

The classical correction for guessing formula of scoring is based on model 1. This formula is, in fact, a *t scale* (scale of tariffs) with special values of TO, TC, and TI presented in Table 1.1. This *t scale* will be referred to as the G *t scale*, where the letter G recalls the words "classical correction for Guessing".

TABLE 1.1. G t Scale (k is the number of alternatives).

In case of	Value of the tariff	Name of the tariff
Omission	0	TO
Correct answer	+ 1	TC
Incorrect answer	$-\frac{1}{k-1}$	TI

The expected score to a question with the G t scale

The *expected score to a question* (ESQ) can be computed with the following general formula:

$$ESQ = (p \cdot TC) + (q \cdot TI)$$

where p = probability of correct answer
q = probability of incorrect answer (q = 1 - p)
TC = tariff in case of correct answer
TI = tariff in case of incorrect answer

The classical correction for guessing formula, that is the G t scale, is conceived in such a way that the ESQ be equal to 0 *in the case of random guessing* (rg) among the k alternatives of a multiple choice question.

This can easily be shown by the general formula if we replace p by the value of p in case of random guessing, that is by $p(rg) = \frac{1}{k}$

and the TC and TI by their values in the G tariff scale, that is by +1 and $\frac{-1}{k-1}$:

$$\begin{aligned} ESQ\ (rg) &= (p(rg) \cdot 1) + (q(rg) \cdot \tfrac{-1}{k-1}) \\ &= \tfrac{1}{k} \cdot 1 + [(1 - \tfrac{1}{k}) \cdot \tfrac{-1}{k-1}] \\ &= \tfrac{1}{k} + [(\tfrac{k}{k} - \tfrac{1}{k}) \cdot \tfrac{-1}{k-1}] \\ &= \tfrac{1}{k} + [\tfrac{k-1}{k} \cdot \tfrac{-1}{k-1}] = \tfrac{1}{k} + [\tfrac{-1}{k}] = 0 \end{aligned}$$

Variations of the expected score to a question (ESQ) according to the values of p.

With the G t scale in effect, it is possible to compute the expected score (ESQ) not only for the random guessing situation (when p = p(rg)) but for each of the possible values of p. The results can be summarized by drawing the function.

When k (number of alternatives) is equal to 4, TI is equal to 0.33 (that is $\frac{1}{k-1}$ and p (rg) is equal to 0.25.The resulting ESQ function (presented in Figure 1.1) is obtained by joining TC to TI. When k is equal to 2 (true-false questions), TI is equal to -1 and p(rg) is equal to 0.5. The resulting ESQ is as follows:

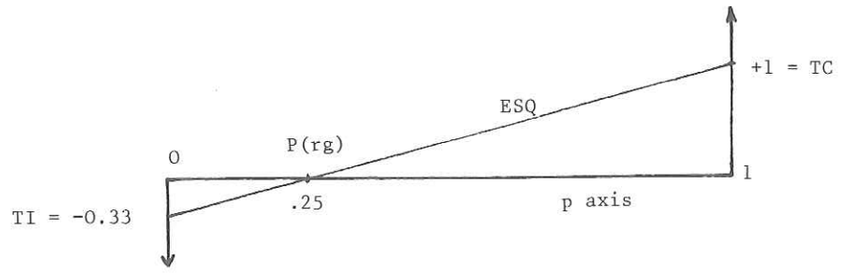


Fig. 1.1. Expected Score to a Question (ESQ) when k = 4 and when the G t Scale is in effect.

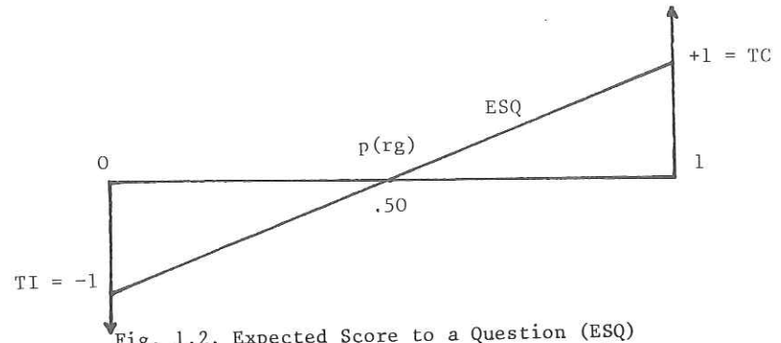


Fig. 1.2. Expected Score to a Question (ESQ) when k = 2 and when the G t scale is in effect.

Criticisms of the classical correction for guessing formula

It is well known that this formula (G t scale) is unfair. In some cases, it undercorrects whereas in other cases it overcorrects.

For instance, it overcorrects since it assumes that each incorrect answer is the result of a random choice (as the model 1 states). It implies that the number of correct answers must be reduced by a portion of *all* these incorrect answers.

This appears clearly in the formula for correcting the total score of a test (ST) on the basis of the student's number of correct responses (NCR) and the number of incorrect responses (NIR). The corrected test score will be referred to as SGT (G meaning guessing).

$$SGT = NCR - \frac{NIR}{k-1}$$

This well known formula is sometimes referred to as the Davis formula, but it was used in 1920 by McCall and was first criticized by West in 1923.

In the following section (devoted to model 2), it will be shown that the correction for guessing based on this first model undercorrects as well.

Alternative approaches

It is obvious that the attractiveness of the various alternatives of a multiple choice question are not equal. Various authors have suggested that the scoring tariffs should be computed according to the statistically observed attractiveness (cf., Chernoff, 1962).

R t scale. Risse (1972) uses the following R t scale:

$$TC = op \cdot oq = op\,(1 - op)$$
$$TI = -(op)^2$$

The R t scale shows a great severity for (statistically) rare errors, where op is the observed proportion of choice of the correct answer. Statistically, the expected score for any question (ESQ) is 0.

F and SD t scales. An equivalent scale (F t scale) has been suggested by Fabre (1975):

$$TC = op \text{ and } TI = 0$$

This last author mentions two other t scales based on the standard deviation of the observed scores to the question (SQ) when TC = +1 and TI = 0. The mean of the SQ is op and the standard deviation is $\sqrt{op.oq}$

The first S D t scale is as follows:

$$TC = 1/\sqrt{op}$$
$$TI = 0$$

The second S D t scale is:

$$TC = \sqrt{oq/op}$$
$$TI = -\sqrt{op/oq}$$

This second S D t scale gives good payoffs to (statistically) rare successes.
W t scale. More sophisticated t scales can be constructed according to the observed popularity of each *distractor* (op) and to the sum of op (that is W). In the W t scale,

$$TC = +1$$
$$TI_i = \frac{W - op_i}{op_i} \quad \text{for the choice of distractor i.}$$

The W t scale punishes the choice of statistically rarely chosen distractors.

Discussion. In spite of the sophistication of such approaches, they are not founded on sound bases for at least two major reasons.

First, popularity (percentage of choices) is not a good index of attractiveness, as the following (theoretical) example shows.

Ten students have to choose among two alternatives. All ten choose alternative 1. They are also requested to express their subjective preferences (see Table 1.2) for each alternative. The average subjective preference for alternative 2 is 40 (on 100) whereas the popularity (pop) is 0.

TABLE 1.2. Subjective probabilities attributed by ten students to each of the two alternatives of a question.

Student	1	2	3	4	5	6	7	8	9	10	Mean	Popularity
Alternative 1	60	55	65	60	55	60	55	65	60	65	60	100
Alternative 2	40	45	35	40	45	40	45	35	40	35	40	0

Secondly, statistical (average) attractiveness is not a good estimation of individual attractiveness. It is obvious that the facility of a question depends upon the student's *personal* ability and this differs from individual to individual. The point of view is supported by recent theoretical work such as in the Rasch model (see Choppin, 1980; Wright & Stone, 1979). Item difficulty must be combined with student's ability to explain the probability of a correct answer.

The obvious way of coping seriously with the problems is to take into account the personal attractiveness, and this will be the job of model 3.

Accordingly, the tariffs depend on the subjective probability (sp) *and* on the correctness of the response. So, we have a *matrix of tariffs* or a *t matrix*.

We shall refer to Educational tariff matrices (or E t matrices) when the TC values are positive, the TI values negative, and TO (tariff of omit) equal to zero.

THE SECOND MODEL OF MENTAL ACTIVITY IN MULTIPLE CHOICE ANSWERING

The Statement of Model 2 (cf., Choppin, 1971)

- When a student knows the correct answer, he chooses the corresponding alternative.
- When he does not know the correct answer, he eliminates the alternatives he knows are incorrect, then chooses randomly among the remaining ones.

In the last case, the student's probablity (p) of getting the correct answer is greater than $1/k$. In fact, it is equal to $\frac{1}{k-e}$ where e refers to the number of eliminated alternatives.

There is no need for deep introspection or numerous interviews to accept that this second model is a more relevant one than the first one. Of course, e is not known in advance, whereas k is unambiguous. It is even reasonable to consider that, in many cases, e is not an integer, but a fraction. This will be taken up in model 3 later.

Model 2 and expected score to a question

According to model 2, the classical correction for guessing (that is, the G t scale) *undercorrects* since $p > 1/k$, so ESQ is greater than 0. The probability is equal to $1/k$ only in limit and in unusual cases. In all other cases, the student is more interested in random guessing among the remaining alternatives than in omitting *even when* the classical correction for guessing (G t scale) is applied.

Frequently, this classical G scale of tariffs (TC = +1, TO = 0, TI = -1/k-1) is presented with the following comment:

"It is in your interest to omit rather than to guess blindly."

Strictly speaking, this comment is wrong because, even with model 1 (relevant only for very few cases), the expected score to the question (ESQ) is *the same* for omission and blind guessing. It is zero. In the majority of the cases (where model 2 is relevant), it is in the interest of the students *not* to omit. This has been shown by using appropriate experimental designs.

Experimental proofs of the superiority of Model 2 vs Model 1

Among various approaches used to study the relevance of omitting when a G t scale is in effect, the most direct one is the so-called "blue and red pencils answering" procedure.

Students are first advised that the G t scale (correction for guessing) will be applied and that, consequently, they should omit when they do not know. During this part of the test, the students answer with a blue pencil. Secondly, the students are requested to answer the questions they had omitted and answer them with a red pencil. The instructions imply that no correction for guessing will be applied to the red answers; only the simple tariff scale (S t scale) will be used.

The average expected score for red responses should be theoretically (according to model 1) equal to 1/k. In fact, the average individual "red score" is largely (and significantly) superior to 1/k.

Cross and Frary (1977) applied this procedure to a test containing 40 multiple choice items with 4 alternatives each. The average expected red score for a question was, of course, 0.25 whereas the average observed red score was

0.328. This last value is close to 0.333 (1/k-1), as if the students had guessed among only three alternatives and not four, since they were able to eliminate one alternative (according to model 2).

Testing and scoring instructions based on Model 2

Coombs, Milholland, and Womer (1956) suggested requesting the student not to choose the correct alternative, but to eliminate the incorrect ones. Thus, faced with a multiple choice item that has only one correct alternative among the k suggested ones, the student can give k - 1 correct answers by eliminating k - 1 incorrect alternatives.

If the tariff for a correct answer is 1 point (TC = +1), k - 1 correct *answers* will result in a score to this question (SQ) equal to k - 1 points. Eliminating no alternative (that is the omission) will result in a score (SQ) equal to 0. It is reasonable to attribute the same 0 score for the question (SQ) to a student that has eliminated all the alternatives, including the correct one.

In this case, the *penalty* (TI) for eliminating the correct alternative must be k - 1.

Generalizing G t scale in case of several correct alternatives

When a multiple choice question contains c correct alternatives among the k suggested ones, the classical correction for guessing formula (G t scale) must be changed: TI (Tariff for incorrect response) is no longer equal to $-\frac{1}{k-1}$ but is now equal to $-\frac{c}{k-c}$.

For instance, if a multiple choice question presents 3 correct alternatives among four incorrect ones (k = 4 and c = 3) in the G t scale, TI must be equal to -3.

Table 1.3 presents the patterns of responses of eight students (A - H) and the scores they obtained for this question (the correct alternative is the second one). Remember that the students must eliminate incorrect alternatives.

TABLE 1.3. Crosses over four Alternatives made by eight Students. The Best Set of Crosses has been given by Student C.

		STUDENTS							
ALTERNATIVES		A	B	C	D	E	F	G	H
Correct one	(1)	1	1	X	1	X	X	X	1
	(2)	2	2	2	2	X	X	X	X
	(3)	X	X	X	3	X	3	3	3
	(4)	4	X	X	4	X	X	4	4
Score for the question (SQ)		1	2	3	0	0	-1	-2	-3

THE THIRD MODEL OF MENTAL ACTIVITY IN MULTIPLE CHOICE ANSWERING

As will be seen below, Model 3 is distinguishable from the two other models by the fact that it does not refer to the dichotomy "when the student *knows*... when the student does *not know*." The third model is not based on a dual representation of knowledge, on two states of mind: nothing *vs* everything. On the contrary, this model supposes a continuity in cognitive stages, from perfect knowledge (e.g., the answer to "2 + 2 = ?") to null knowledge (e.g., the answer to "which day of the week was Einstein born?"). *Between these two extremes* exist a lot of intermediate states of knowledge, what is often referred to by the expression *partial information*. For instance, a student may not know exactly in which country the town Quito is, but is able to eliminate countries like the UK, USSR, USA, etc.

Statement of Model 3 (Cf. Choppin, 1971)

> When a student is faced with the various alternatives, he attributes to each of them a probablity of being the correct one. Since he is requested to give only one answer, he will choose the alternative with the greatest probability.

It must be noted that the word probabilty is a theoretical way of describing the cognitive process. Actually, most students *rank* the alternatives from the most likely one to the least likely one, without attributing numeric values to these likelihoods. But when the instructions request the student to express these likelihoods with numbers, most students are able to do so.

Is it reasonable to ask students to answer in this way? More and more researchers and teachers answer yes, sharing De Finetti's famous options (1965, p. 109):

> Partial information exists; to detect it is interesting, necessary and feasible.
>
> Instruction in using the methods with which we are concerned has, moreover, a high educational value.
>
> Such methods, INCLUDING THE WAY OF SCORING, and not only the response systems, must be appropriately chosen by the experimenter and clearly explained to the subjects who must understand the nature of the game they are playing.
>
> If this is done, questions about guessing disappear.

Three basic concepts in decision making

In decision theory, the central concepts are the possible acts, the states of nature and the consequences on payoffs (here tariffs).

According to Lindley (1971, p. 4) a decision can be defined as choosing one of the possible acts. This choice depends only on the student.

The state of nature is the unknown event. For instance, when meteorologists have to forecast on Monday what the weather will be on the next day, the state of nature is the observed weather on Tuesday (sunshine, cloudy, rain, snow, hail). The decision maker has to consider future situations over which he is powerless. Hesitation in choosing which actions to take comes in part from uncertainty about the (future) states of nature.

The matrix of tariffs is the table of tariffs associated with choice of actions A1, A2, ...when the states of nature are SN1, SN2, etc. respectively. In the weather forecasting game, the payoffs could be positive, null, and negative points, as in the following Matrix A (see Table 1.4). In the complete matrix A, consequences are fixed (+2) when the prediction is verified (diagonal of the matrix), but consequences vary according to the type of discrepancy in the case of failure (from -1 to -4).

Often, decision situations give a fixed payoff for a correct answer and another fixed one for an incorrect answer so that only two states of nature (or events) have to be considered: success or error. Matrix B contains tariffs values of such a payoff matrix.

TABLE 1.4. Two Theoretical Matrixes of Tariffs for Weather Forecasting.

Matrix A

Decisions (forecasts) \ (Future) States of Nature	SN1 Sunshine	SN2 Clouds	SN3 Rain	SN4 Snow	SN5 Hail
A1 Sunshine	+2	-1	-2	-3	-4
A2 Clouds	-1	+2	-1	-2	-3
A3 Rain	-2	-1	+2	-1	-2
A4 Snow	-3	-2	-1	+2	-1
A5 Hail	-4	-3	-2	-1	+2

Matrix B

Decisions (forecasts) \ Future event	Correct	Incorrect
A1 Sunshine	+2	-3
A2 Clouds	+2	-1
A3 Rain	+2	-1
A4 Snow	+2	-2
A5 Hail	+2	-2

The evolution of decision theory

The starting point: (objective) probabilities and (objective) values. Until the eighteenth century, the three above mentioned concepts (possible acts, states of nature, and payoffs) were the bases of decision theory. The principal formula to compute the *expected value* (EV) of the score at a question (SQ) was as follows:

$$EVSQ = (p \cdot TC) + (q \cdot TI)$$

where
- p = (objective) probability of success.
- TC = The value of the tariff (or payoff) for a correct answer.
- q = (objective) probablity of failure.
- TI = the value of the tariff (or payoff) for an incorrect answer.

Utility (or subjective values). In 1738, Bernouilli suggested substituting the concept of utility for the concept of value. Whereas there is an objective scale of values, the utilities are essentially subjective; they vary from one individual to another. For instance, a banknote of one English pound has the *same value* (of one pound) for everybody, whereas it does not have the *same utility* (it is more appreciated by poor people than rich ones).

"Utility is a number that measures the attractiveness of a consequence - the greater the utility, the more desirable the consequence - the measure being done on a probability scale." (Lindley, 1971, p. 70)

If all the utilities are multiplied by a same positive constant, and if we add to them any other constant, the resulting numbers lead to the same decisions as the original ones, so there are many equivalent utility scales.

Utility functions. Figure 1.3 presents the utilities plotted against the objective values (expressed in US dollars). The result is a concave increasing ogive with a horizontal asymptote. Letters A, B, C, D, P, Q, R, and S have been added to the diagram in order to stress that, whereas the distance from A to B is equal to the distance from C to D, the increase of utility from R to S is far less than the increase of utility from P to Q.

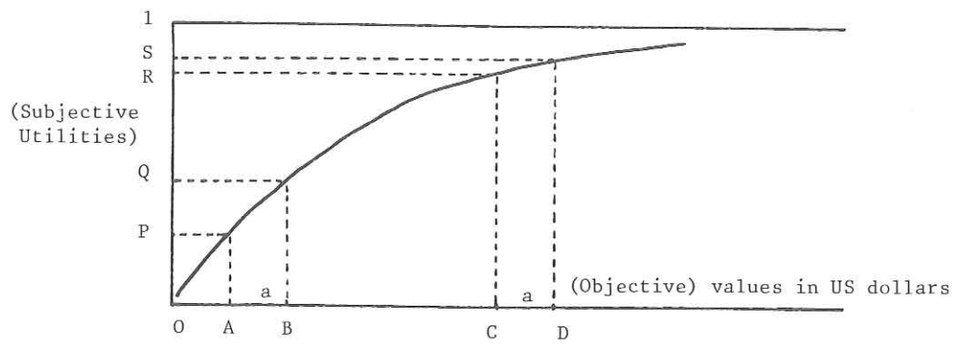


Fig. 1.3. An individual's utility curve for receiving increasing amounts of money.

Utility and logarithmic functions in psychophysics. At the end of the 18th century, it has been supported that "Every man's utility function" was defined as follows:

$$u = \sqrt{v} \text{ if } v \geqslant 0$$
$$u = -(v^2) \text{ if } v \leqslant 0$$

Graphically, it is shown in Figure 1.4. The logarithmic aspect of utility functions influenced deeply the work of Fechner and the field of psychophysics.

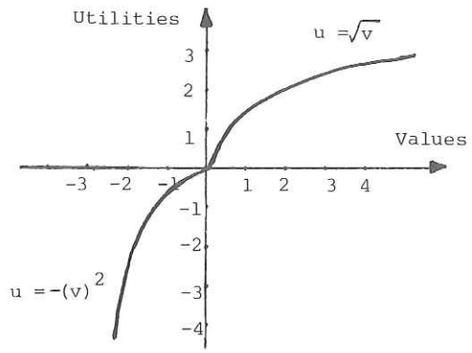


Fig. 1.4. Every man's utility function.

Is the Utility curve of examination scores an ogive? Whereas this problem is central in confidence marking, we know only one attempt to answer this question on experimental bases. Van Naerssen, Sandbergen, and Bruyniss (1966) administered a test to students where the payoff depended not only on the number of correct answers, but moreover, on a confidence index Z (*zekerheid* means *certainty* in Dutch). When the students are confident about their answer, they add a Z next to it.

Their fundamental hypothesis was as follows: "Students with a low number of correct items are anxious about their results, so they are expected to give relatively more Z than students with a great number of correct items," for items of equal objective facility (p value) for the groups, respectively A (high scores), B (medium scores) and C (low scores).

Their procedure was as follows: On the basis of the p indexes, pairs of items have been constituted so that one of the two items has the same rate of success in group A as the other item has in group B. For instance, in the 1964 experiment:

39 pairs of items could be constituted for groups A and B.
34 pairs of items for groups B and C.
26 pairs of items fof groups A and C.

In each of those three series of items, three groups of items have been constituted according to the difficulty.

1 = difficult items.
2 = intermediate items.
3 = easy items.

For A1, A2, A3, B1, B2, B3, C1, C2, and C3, a $\bar{p}$ (average value of p, that is, rate of success, and a $\bar{z}$ (average rate of use of Z index) have been computed.

Those nine values of $\bar{p}$ and nine values of $\bar{z}$ have been plotted in the graph presented in Figure 1.5. The whole procedure was repeated in 1965. The two curves display an inverted S shape (of which the left side is invisible by

shortage of difficult items. If the hypothesis of ogiveness of the utility function were true, B points would be on the curve, A ones under the curve, and C ones over the curve. Neither the graph, nor the tests of significance are in favor of this hypothesis.

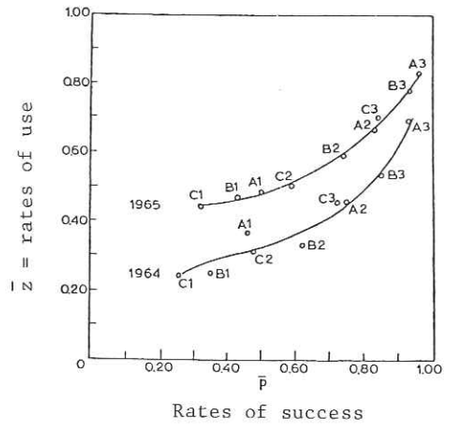


Fig. 1.5. Relationship between the $\bar{z}$ (average rate of use) and $\bar{p}$ (average rate of success) values for three groups of items (1, 2 and 3) answered by the three groups of students (A, B and C).

On the basis of this single (complicated) experiment, we shall go on as if the utility function were linear, that is, as if utilities and values could be confounded.

Nevertheless, it is obvious that more experimental light is needed on this point, where time for conclusion has not yet come.

The utility equation. When utility is considered, the basic question is modified.

$$EVSQ = (p \,.\, TC) + (q \,.\, TI)$$

becomes

$$E\,U_i\ SQ = (p \,.\, U_i\,TC) + (q \,.\, U_i\,TI)$$

where EU_i = expected utility for individual i
U_i TC = utiility for individual i of tariff TC
U_i TI = utility for individual i of tariff TI

Since Van Naerssen, Sandbergen and Bruyniss' results did not support the hypothesis of curvilinearity of utilities, what follows below will be developed assuming that utilities (U) and values (V) of tariffs (TC and TI) can be confounded, i.e., that their relation is linear and not curvilinear.

So, for examination scores, the expected utility formula will be written:

$$EU_i\,SQ = (p \,.\, TC) + (q \,.\, TI)$$

without the U_i coefficients for the TC and TI values.

Subjective probability. In the middle of the twentieth century, subjective probabilities were substituted for objective ones. The leaders of this movement were Savage (1954), De Finetti (1965, 1971), and Raiffa (1971).

The basic formula now includes the subjectively expected (S_iE) score for a question (SQ). The letter S is written with an i subscript to indicate that the subjective estimation is made by the individual i.

$$S_iESQ = (sp_i \,.\, TC) + (sq_i \,.\, TI)$$

where S_iE SQ is, for individual i, the subjectively exptected score to a question.
sp_i is the subjective probability of success, estimated by individual i.
sq_i is the subjective probability of failure, estimated by individual i.

Modern utility theory. According to the "modern utility theory" (von Neumann & Morgenstern, 1947), students should behave to maximize this S_iESQ value for each question. Do they behave in this way? In fact, as will be seen in Chapter 5, faced for the first time by a confidence marking procedure, the students generally use a lot of inefficient (but attractive) strategies. Nevertheless, students should be trained in answering with confidence degrees since this is the only theoretically well founded way of responding. As De Finetti (1965, p. 110) points out: "Instructions like 'mark an alternative only if you KNOW it is right' or 'cross it out only if you KNOW it is wrong' are unavoidably ambiguous just because of their apparent precision which is so absolute as to be illusory."

DEVELOPING A NEW MODEL OF EDUCATIONAL MEASUREMENT

From Miller (1956), it is well known that, without specific training, the capacity of human beings in perceptual activities is restricted to "The magical number seven, plus or minus two." Whereas human limitations of acuity in self estimation should be carefully measured (see Chapter 6), they undoubtedly exist and affect the transformation of ability to esp (see Figure 1.6).

Fig. 1.6. Representation of human average acuity as an intervening variable in the transformation of ability into the estimated subjective probability (esp.).

Can the experimenter trust the *observed subjective probabilities* (osp) provided by the student? Stated in other words, is the osp an unbiased reflection of esp and then, of ability? (See Figure 1.7.)

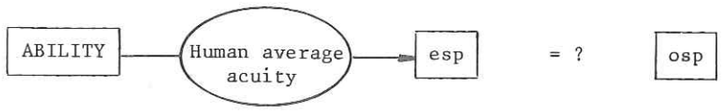


Fig. 1.7. Representation of the problem of equivalence between osp and esp.

The answer is negative if a lot of (external and internal) intervening variables are not taken into account:
- test instructions
- individual realism in self assessment
- the tariffs and attitude towards risk (personality)
- familiarity with the process (training).

The test instructions

Let us consider that the student has reached the conclusion that, for a given answer, his estimated subjective probability (esp) value varies from .5 to .6. If the instructions say "Just tell me whether you are sure or not about your answer", the student will make a decision according to his personal interpretation of the word "sure", and the observed expression of confidence (osp) will be far less accurate than the estimated subjective probability (esp). The place of instructions in the process of transforming esp into osp is presented in Figure 1.8.

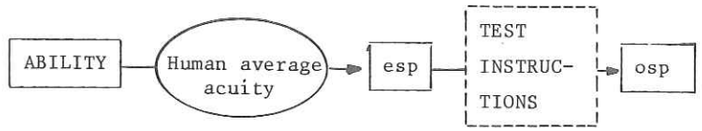


Fig. 1.8. Test instructions presented as intervening variables in the transformation of esp into observable subjective probability (osp).

The individual's realism in self assessment

Individuals differ in their capacity of self assessment, that is, in their realism. It is well known that some systematically overestimate themselves whereas others are underestimators. This problem will be deepened in Chapter 4. The realism of self assessment appears as an intervening variable in Figure 1.9.

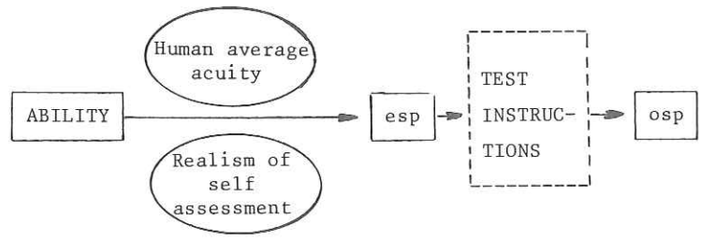


Fig. 1.9. Realism of selfassessment as an intervening variable in the transformation of the ability into the estimated subjective probability.

Training can affect positively both the individual's average acuity and his realism in self assessment (see Figure 1.10).

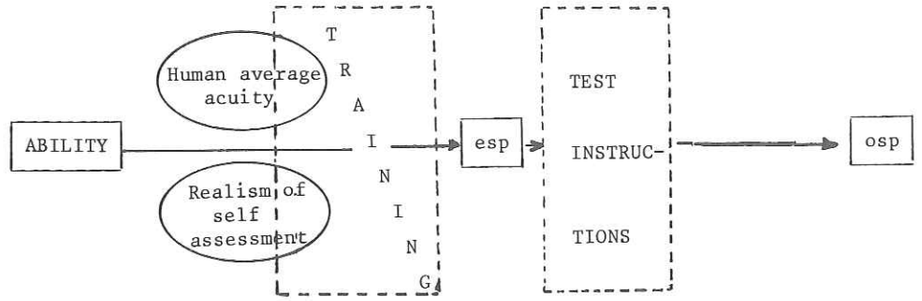


Fig. 1.10. Training, an intervening variable improving two other intervening variables (acuity and realism).

The transformation of the estimated subjective probabilities (esp) into observable subjective probabilities (osp) is mainly influenced by the student's attitude toward risk.

The individual's attitude towards risk

As will be seen in Chapter 4, with many scales of tariffs, the way to maximize one's expected score to a question (ESQ) is not by telling the truth, but by choosing a given probabilty (osp) different from one's esp (see Figure 1.11).

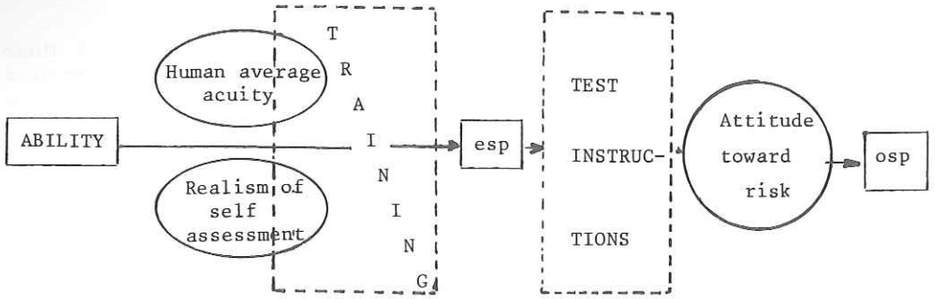


Fig. 1.11. The individual's attitude toward risk as an intervening variable in the tranformation of esp into osp.

This problem is exemplified by the differences between "pure strategies" and "matching strategies".

Suppose that a physician is faced with an epidemie that attacked 1000 people. Three drugs are available: X, Y, and Z.
Statistically, X results in recovery in 70% of cases.
Y results in recovery in 40% of cases.
Z results in recovery in 20% of cases.

The doctor has each of the drugs in sufficient quantities to treat the whole population, but the drugs are exclusive of each other: A patient may not take two different drugs. What should the doctor do? Of course, he should give drug X to all the patients because the expected number of recoveries is .7 x 1000 = 700 persons. Any other strategy would result in a lower expected number of recoveries. It appears that the rate of use (RU), here 100%, depends upon the rate of success (RS) but that these two measures are distinct.

The doctor has used a *pure strategy*. This was predictable because the doctor's objective is to maximize the number of successes (here, recoveries). This maximization is not always the goal of the decision maker. The following experiment is an example of a situation where *matching strategy* is adopted.

Siegel (1961) presented to four year-old children two opaque bottles, one of the two containing a reward. At each trial, the child indicated one of the two bottles and the experimenter turned it upside down. If a reward appeared, the child kept it: it belonged to him. A random sequence was executed by the experimenter. In this sequence, the reward was placed in 75% of the cases into the left bottle and in 25% of the cases into the right one.

At the beginning, the children chose the two bottles with about the same frequency, but after a few dozen trials, children indicate unvaryingly the left one (pure strategy).

Siegel and McMichael (1961) presented the same game to equivalent children but they changed the nature of the reward. In the previous experiment, it was a little toy, varying with each trial (a little car, poker dice, marble, etc.). In the second experiment, the reward was always the same (a collar-stud). In this case, children adopted a matching strategy. In the long run, they indicated 75 times the left bottle and 25 times the right one in 100 trials.

This change of behavior can easily be explained. Since no reward (psychologists would say "positive reinforcement") is associated with success, individuals adopt a more varied behavior, a less boring one (remember that the pure strategy implies always the same action: pointing to the left bottle). The second experiment appears more as a game (where wins are purely psychological but actions are varied) whereas the first experiment appears as work (where salary is more tangible but the actions are less attractive). This implies that in confidence testing, tariffs (see Figure 1.12) must be carefully chosen and computed in order to favor students who tell the truth. (This problem will be examined in Chapter 4.)

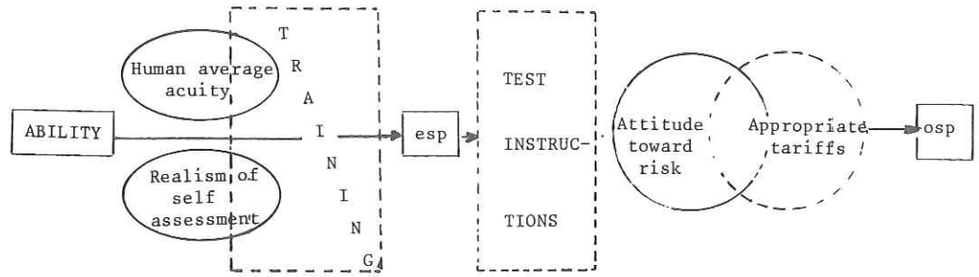


Fig. 1.12. Appropriate tariffs: an intervening variable improving another intervening variable (attitude toward risk).

Even when instructions and tariffs are constructed so as to avoid biasing esp, students often do not understand that the best strategy is to tell the truth. For this reason, training should be undertaken in order to make the tariffs more familiar to the student (Figure 1.13).

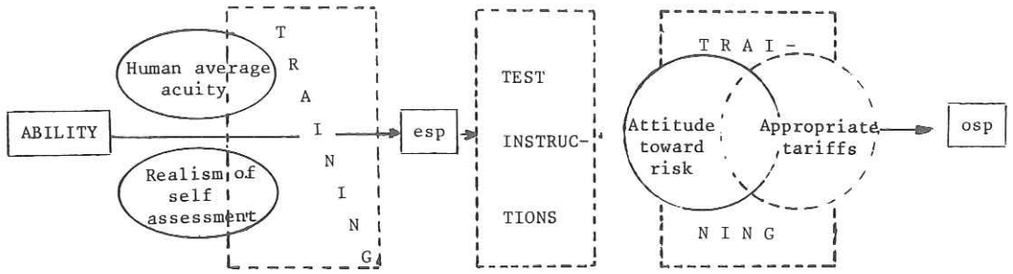


Fig. 1.13. Training : an intervening variable improving another intervening variable (attitude toward risk).

Taking up a challenge

The objection may be made that what the experimenter observes (the osp) and what he is interested in (ability or objective knowledge) are so far from each other (see Figure 1.13) that confidence marking is of low interest. There are many reasons for rejecting this pessimistic point of view.

- Reality is complex and the best way to understand it is to point out clearly where the intervening variables are, whatever their number, and to cope with the situation, however complicated it may be. What is surely undesirable is to try to handle a complex situation as if it were simple.

- Even in classical approaches to the correction for guessing, researchers have introduced personality as a factor. For instance, Ziller (1957) and Slakter (1968) have proposed indexes to estimate the individual's tendency to guess. Corresponding corrections-for-guessing are derived from these indexes.

- It is possible to clear up these problems.

First, it is possible to reduce the importance of "the attitude towards risk" by computing payoff matrices (i.e., tariffs) in order that students maximize their expected score *by telling the truth*, that is, by expressing their estimation of subjective probability without faking. Second, it is possible, with simple concepts and formulas, to estimate the individual's capacity of self assessment, then to conceive of ways to improve it. Third, it is possible to train students to understand the instructions.

In various experiments or theories (some of them will be cited hereafter), the whole approach is invalidated because one or several of the above variables have not been considered. Only one wrong concept or one bad procedure or an inappropriate t scale is sufficient to produce useless data. To cope with the complexity of the problem, attention must be paid simultaneously to all aspects. When those precautions are taken, a promising landscape appears for research and practice.

THE SOCIAL BENEFIT OF THE CONFIDENCE APPROACH

School sometimes penalizes omission or doubt; consequently, it reinforces unwarranted pretentions of confidence. For instance, in oral examination, if a student doesn't know the answer to a question, he sometimes prefers to fill the gap with "neutral or vague" considerations rather than to confess his ignorance or uncertainty. The problem is that, in many occasions, the teacher's tariffs reinforce this way of answering.

Such an avoidance of clear appraising of truth is not a sane preparation to adult responsabilities. Think, for example, of the importance of a clear estimation of chances of success for physicians, nurses, and surgeons in a hospital, for teachers in a school, workers in a factory or drivers on the roads. If someone is strongly sure of what he says and is wrong half of the time, he is a perpetual source of annoyance and danger to himself and others. It is also crucial that adults estimate the chances of success in any project they undertake.

When someone has doubts about the spelling of a word, normal behavior does not consist of audaciously trying the best guess but of consulting a dictionary. Why should such behavior be humiliating ? The more typical characteristic of a "sensible" person is that he knows what he does not know. As noted by Russel, the problem in our world is that foolish people are sure of themselves whereas sensible people are filled with doubt.

Anyone should be open to criticism, able to use efficiently others' judgments without loosing self esteem, i.e., without excessive vulnerability but, at the same time, without indifference. School has a great responsibility for developing these cognitive and emotional mechanisms.

CHAPTER 2

INSTRUCTIONS AND TARIFFS IN CONFIDENCE MARKING

THE VARIOUS WAYS OF DEFINING THE WORD CONFIDENCE

Many experimenters request students to express their confidence by choosing among the degrees on an ordinal scale. That kind of instruction is irrelevant as shall be seen. In the following examples, some payoff matrices will accompany the instructions whereas the problem of the consequences (tariffs) will be treated only in the next section.

Lay Stress on Your Answer

This kind of instruction reads as follows:

> Underline (or circle or mark by a star) your answer if you are sure of it.

Van Naerssen and Van Beaumont (1965) type their multiple choice items as follows:

> What is the capital of France?
> Z 1 - Lyon
> Z 2 - Paris
> Z 3 - Marseille

Z is the first letter of *zekerheid* ("certainty" in Dutch). The students can choose a solution (2 for instance) either by circling
only 2 (not confident : z (2))
or by circling Z and 2 (confident : (z 2))

Their tariff matrix appears in Table 2.1.

Sandbergen (1971) used the same matrix as Van Naerssen and Van Beaumont except that all the tariffs were doubled.

TABLE 2.1. Van Naerssen and Van Beaumont's Tariff Matrix (1965).

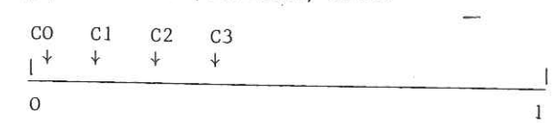
	TC	TI
without Z	+ 0,5	0
with Z	+ 1	- 0,5

Confidence Degrees on Ordinal Scales

These instructions often read as follows:

If you are not sure at all, indicate 0.
If you are weakly sure, indicate 1.
If you are fairly sure, indicate 2.
If you are strongly sure, indicate 3.

Such definitions are vague; what is strongly sure for one student might be fairly sure for another one and weakly sure for a third one. Some over confident students would place such verbal notions (denoted C0, C1, C2, and C3) at the following points on a probability scale:



Other (unconfident) students might place them in other positions on the same probability scale:

Since the experimenter does not know which exact interpretation of "fairly" sure has been given by various students, two answers with confidence degree C2 may not be compared if they are given by two different students. Moreover, nothing prevents the student from changing his interpretation of "strongly sure" during the course of a test. Contrasts between successive questions could lead him to modify his interpretation as he goes from item to item. So two degrees of confidence C2 are hardly comparable even if they are given by the same student.

Notwithstanding those weaknesses, this type of instruction has been profusely used with various tariff matrices. For example, Jacobs (1972) used the tariff matrix presented in Table 2.2.

TABLE 2.2. Jacob's Tariff Matrix (1972)

	TC	TI
I guess	+ 1	0
Fairly sure	+ 2	- 2
Confident	+ 3	- 3

In Belgium, from 1970 (see Leclercq, 1973), the same kind of instructions were used. The tariff matrix was slightly different (see Table 2.3) since omission received a 0 score.

TABLE 2.3. Typical Tariff Matrix (has been used in the 1971 experiment).

	0	25	50	75	100
TCs	0	+1	+2	+3	
TIs	0	-1	-2	-3	

This tariff matrix appeared to be convenient for practical purposes: values are easy to recall since they are equal to the codes of the confidence degrees and the hand computation of the test score is easy. Nevertheless, for reasons that will be developed in Chapter 3, this matrix will be abandoned since it does not respect some criteria from decision theory.

Confidence Degrees on Regular Zones of Interval Scales

For reasons that have been explained above, the expression of confidence should be made in probabilistic terms. Nevertheless, it is difficult for a student to distinguish between being confident at the .372 level and the .373 level. Indeed, we know ourselves that it is difficult to distinguish .3 from .35. So, it would appear useless to request answers of such acuity. That is the reason why a few areas are delimited on the probability scale, so that the student has to choose a zone of probabilities and not a pin-point probability.

One of the most common forms of instructions used in Belgium (Leclercq, 1973) is presented in Table 2.4. The payoff matrix on the right was used until 1972. After this period, it was replaced by tariffs computed according to decision theory (see below).

TABLE 2.4. Instructions used from 1971 in Belgium, and Tariff used from 1970 to 1972.

If you attribute to your answer a chance of success	TC	TI
- varying from 0 to 25 %, then choose confidence degree 0.	0	0
- varying from 25 to 50 %, then choose confidence degree 1.	+1	-1
- varying from 50 to 75 %, then choose confidence degree 2.	+2	-2
- varying from 75 to 100 %, then choose confidence degree 3.	+3	-3

Confidence Degrees as a Number of Fixed Amounts of Probabilities

De Finetti (1965) describes a "five star system" where the student has to distribute five stars over the alternatives. Each of the stars is equivalent to a .20 probability value.

With five alternatives, there are only seven ways of distributing five stars:

5	3 - 2
4 - 1	2 - 2 - 1
3 - 1 - 1	1 - 1 - 1 - 1 - 1
2 - 1 - 1 - 1	

Michael (1968) has used a "ten star system" where the score is the number of stars attributed to the correct alternative, number divided by 10. From his experimental data, he concluded that a classically corrected test should be 1.7 times longer to reach the reliability obtained with a ten star system.

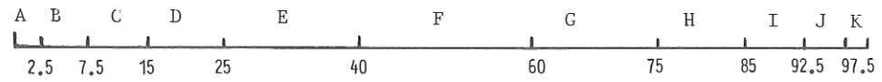
Confidence Degrees as Ratios

Edwards (1968), among others, recommended using a logarithmic scale for defining degrees. Here is an example of such a scale:

A = 1/1000 chance
B = 1/100 chance
C = 1/10 chance
D = 1/4 chance
E = 1/2 chance
F = 3/4 chances
G = 9/10 chances
H = 99/100 chances
I = 999/1000 chances

This scale (9 degrees) is related to the logarithmic properties of human perception. It is likely that Miller's magical number seven (1956) applies to subjective probabilities as well as to visual, tactile, auditive, etc. perceptions. In Chapter 6 (on acuity of confidence weighting), this problem will be examined further. Nevertheless, it is reasonable to suspect that an average student can make a distinction between say, 1/100 chance, and 1/10 chance (a difference of 9% on a percentage scale) whereas he can hardly make a distinction between 33% and 42% (the same difference of 9% but at the middle of the percentage scale).

Bearing this in mind, it is possible to define unequal (but symmetric) zones of confidence on an interval scale. An example is presented in Figure 2.1.



A = K = 2.5 %
B = J = 5 %
C = I = 7.5 %
D = H = 10 %
E = G = 15 %
F = 20 %

Fig. 2.1. Probability scale where unequal (but symetric) zones of confidence have been fixed according to a logarithmic progression.

Confidence Degrees on Irregular and Dissymetric Zones of an Interval Scale

In educational measurement, the areas on the right hand side of the scale are far more used than the areas on the left. This is explained by the fact that tests are mostly used as post-tests rather than pretests to test whether students have mastered the objectives they have just been taught.

For this reason, it is sensible to conceive a scale with the following general format presented in Figure 2.2.

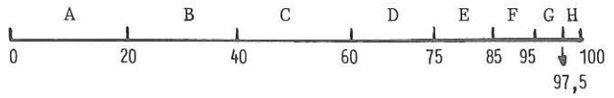


Fig. 2.2. Probability scale with unsymetric and unequal zones fixed according to the general principle of a progression of the logarithmic type.

The Continuous Confidence Marking

The continuous confidence marking procedure, recommended by Di Finetti (1965) allows the student to express his confidence with any precision he likes (.3 as well as .3027). This way of answering is useful when continuity is possible or interesting. Baker (1965), for example presents on a (computer) visual screen the four histograms corresponding to the probabilities (expressed in percentages) attributed to each of the four alternatives of a multiple choice question. This device is presented in Figure 2.3.

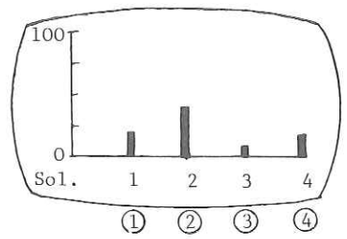


Fig. 2.3. Baker's device, including screen and knobs.

Four knobs enable the student to change each value (the change is immediately displayed on the screen). The three other values are then automatically modified (by a proportionality rule), so that the four values always sum up to 100. In this example, changes need not be done step by step, but may be continuous.

From the scoring point of view, the payoff matrix need not be a set of values but may be a continuous function. Examples of such functions will be given in the following section.

Fractiles

When the response concerns a numeric value of a continuum (for instance, weights, surfaces, prices, dates, durations, speeds, etc.), the most direct way of coping with partial knowledge is the fractile procedure. The student is requested to provide, for each question, two responses (that is, two fractiles): the lower limit and the upper limit of an interval. It is obvious that the narrower the interval, the more confident the student and the greater the risk of the correct answer falling outside the interval...so the higher must be the positive tariff in the case of success and the greater the negative one in the case of failure.

When the interval does not include the correct value, it is referred to as a "surprise". An unpredicted high number of surprises is often observed. Whereas the instructions request the students to give an interval in order that it includes the correct response with a probability of, for instance, .8, a rate of success inferior to .5 is often observed. This phenomenon has been called *interval hyperprecision* by Pitz (1974).

More realistic instructions request the student to give an interval in such a way that one third of the correct answers fall under, one third over, and one third within the interval; this procedure is called "tertiles" (Pitz, 1974). Alpert and Raiffa (1969) use five fractiles, Shaefer and Borcherding (1973) use seven.

The whole problem has been comprehensively reviewed in Lichtenstein et al. (1977), Murphy and Winkler (1974), and Hardy (1981).

Discussion

Only the above procedures numbered from 3 to 8 can be used in what Shufford, Albert and Massengill (1966) call "admissible probability measurement procedures".

The choice between procedures 3 to 8 depends on the purpose of the research or the practice, on the available means, on the degree of familiarity of the student...and of the teacher.

For instance, it is useless to allow the student to give a detailed confidence response (for example, the continuous confidence marking procedure) if the teacher has no means to handle it properly and to provide relevant feedback. If a .372 confidence degree is treated as if it were a .3, the accurate expression of confidence is a loss of energy.

Handling an accurate probability implies the use of appropriate formulas (see below) and computation tools as computer programs.

Young children or untrained students should be given instructions with low number of confidence degrees. The scale presented in Figure 2.4 was revealed to be convenient and well accepted even by ten year old children. If an experimenter is interested in measuring slight changes in knowledge, he will use a greater number of confidence degrees.

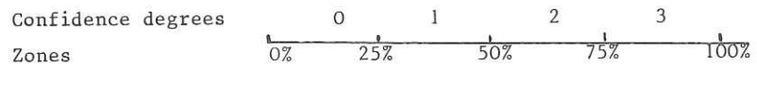


Fig. 2.4. Graphic representation of verbal instructions presented in TABLE 2.4.

THE VARIOUS TYPES OF TARIFF MATRICES

In order to keep the reasoning simple, various possible types of payoff matrices will be illustrated using the instructions presented in Table 2.4 and summarized in Figure 2.4. Various categories of payoff matrices will be considered: from the least structured to the most elaborate. They will then be presented in a coherent model.

A Matrices (A means ANY)

In A matrices, there is no rule for deciding the values of payoffs. Such matrices are useless for educational purposes.

M Matrices (M means MONOTONOUS)

In M matrices, the best tariff is attributed to a correct answer with the greatest degree of confidence, and the worst tariff is attributed to an incorrect answer with the greatest degree of confidence. The tariffs of intermediate degrees can be ranked. An example is presented in Table 2.5.

TABLE 2.5. Arbitrary Example of M Tariff Matrix

Confidence degree	0	1	2	3
TC s	-10	-9	-8	-7
TI s	-15	-16	-17	-23

It must be noted that in such matrices, all the tariffs can be positive or all negative (as in the example above). For this reason, such a matrix is not of great use in educational settings where it is unusual to give points to students who provide a wrong answer or to withdraw points when the correct answer is given.

E Matrices (E means EDUCATIONAL)

In E matrices, the tariffs not only increase monotonically as the confidence degrees do, but in addition,

- tariffs for correct answers are positive,
- tariffs for incorrect answers are negative, and
- tariffs for omission are null.

An example is provided in Table 2.6.

TABLE 2.6. Arbitrary Example of E Tariff Matrix

Degree of confidence	0	1	2	3
TC s	0	+2	+4	+6
TIs	0	-1	-2	-3

Unfortunately, the great majority of such matrices are not compatible with decision theory as will be demonstrated later.

Examining an Educational Payoff Matrix with Decision Theory

Let us consider the educational payoff matrix above. For each degree of confidence (0, 1, 2, and 3), it is possible to draw a function of the expected scores related to the probability of success. Let us recall the formula:

$$\text{For degree 1, } SESQ_1 = (sp \cdot TC_1) + (sq \cdot TI_1)$$

$$\text{In our example, } SESQ_1 = (sp \cdot 2) + (sq \cdot -1)$$

Graphically, SESQ1 (subjectively expected score to the question with a confidence index 1) can be represented as follows by joining TI1 and TC1.

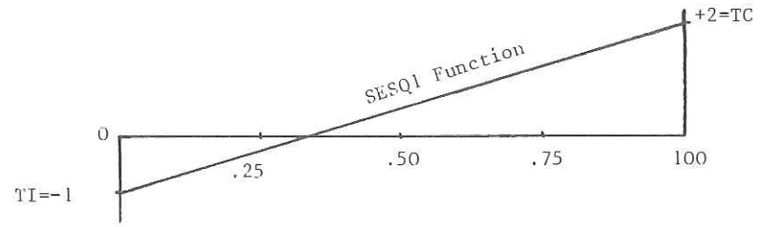


Fig. 2.5. SESQ function of Confidence degree 1 according to tariff matrix presented in TABLE 2.6.

If a student has sp equal to 0.20, his sq is consequently equal to 0.80 and his subjectively expected score if he uses one degree of confidence is:

$$SESQ1 = (0.20 \cdot 2) + (0.80 \cdot -1) = 0.4 - 0.8 = -0.4$$

This case is illustrated by the dotted lines in Figure 2.6.

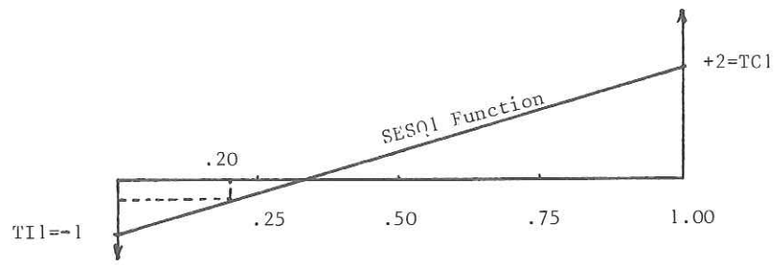


Fig. 2.6. SESQ value giving confidence degree 1 with an esp equal to .20

SESQ2 function may be drawn by joining TI2 (in our example, TI2 = -2) to TC2 (in our example, TC2 = +4).

SESQ3 function may be drawn by joining -3 (TI3) to +6 (TC3). The resulting graph is presented in Figure 2.7.

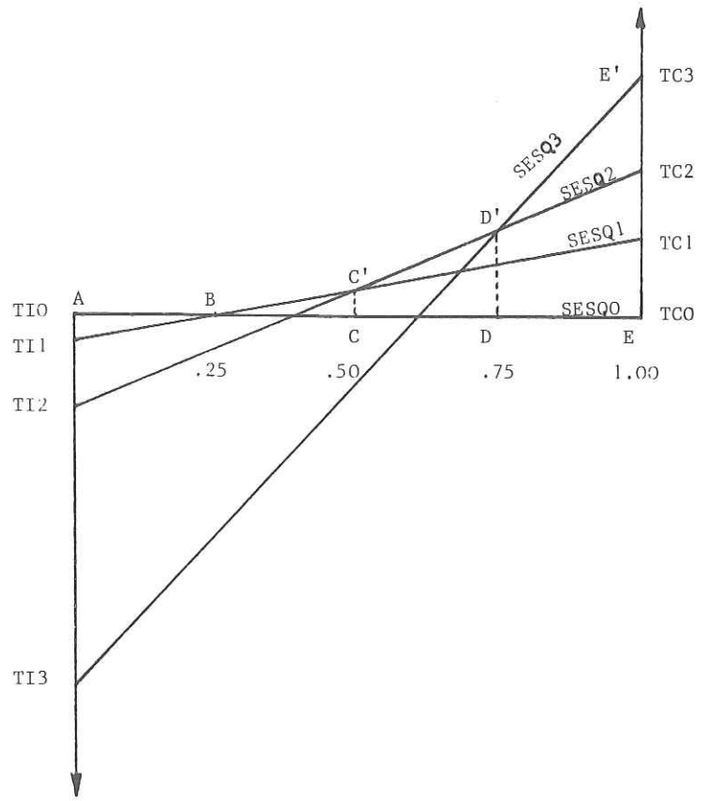


Fig. 2.9. SESQs conceived according to decision theory to accompany the instructions presented in figures 2.4. and 2.8.

Determining Starting Values

The TC values may be fixed arbitrarily, and the TI will follow mathematically from this choice.

An example is presented in Figure 2.10. TC2 has been fixed at +2; TC1 has been fixed at 0.25 below (that is, at + 1.75) and TC3 at 0.25 above (that is at +2.25). The resulting TI1, TI2 and TI3 were respectively equal to - 0.53 (TI1), - 0.83 (TI2), and - 1.57 (TI3).
This is a D matrix (consistent with decision theory). There is an infinity of such matrices. They can be defined by fixing the TCs or the TIs. For educational purposes, it is interesting to focus on D matrices that would present only *integer* payoffs, that is on I matrices.

The Computation of I Matrices

When the number of areas of confidence on the probability axis is small (4 or 5), it is possible to derive such I matrices, with a ruler, by trial and error.

For instance, the I matrix presented in Figure 2.11 has been derived by trial and error using graphs.

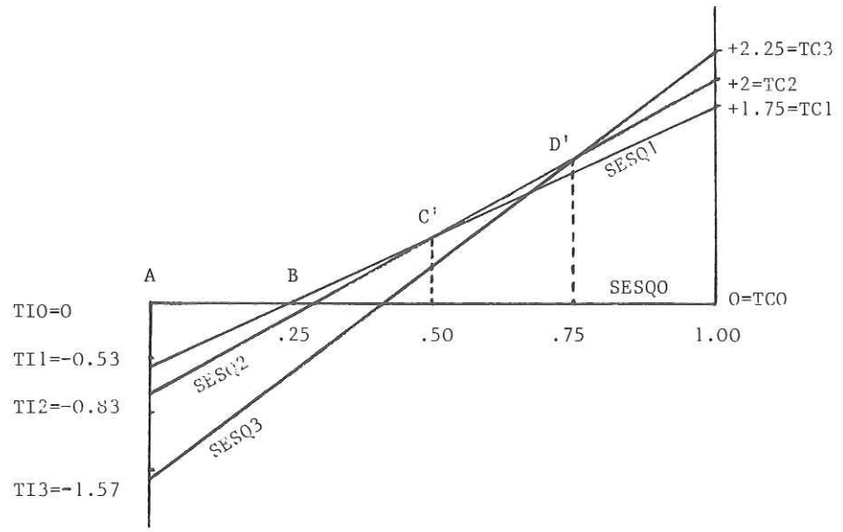


Fig. 2.10. An example of DESQs (compatible with decision theory and instructions presented in figures 2.4. and 2.8), where the TCs have been arbitrary choosen.

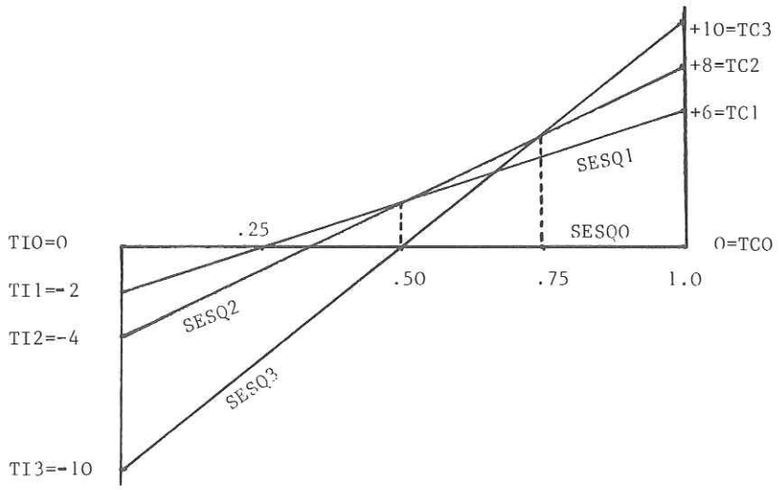


Fig. 2.11. Relations of inclusions between the various kinds of tariff matrixes.

It appears that if all the tariffs were multiplied by any constant, the resulting tariffs would still constitute a D matrix. Especially, if these tariffs are timed by 0.5, it will result in another I matrix (see Table 2.7). This last I matrix is a very special one: it is the D matrix for those limits on the axis (.25, .50, .75, 1), that present *the least integers*. That is the reason why such matrices are called *L Matrices*.

TABLE 2.7. Typical L Matrix.

p	0-25	25-50	50-75	75-100
TC	0	+3	+4	+5
TI	0	-1	-2	-5

A Model of Classification of Tariff Matrices

The relations between all those matrices, i.e., between monotonous (M), educational (E), decision theory (D), integer (I), and least integer (L) matrices are represented in Figure 2.12.

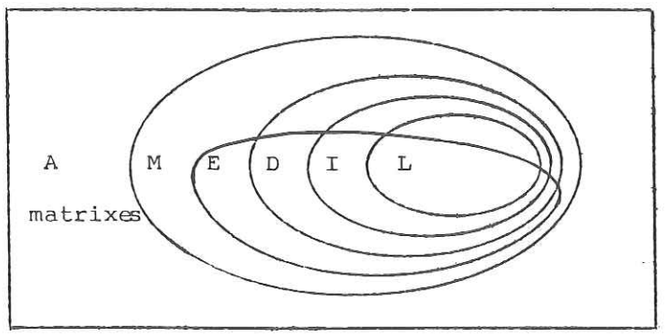


Fig. 2.12. Relations of inclusions between the various kinds of tariff matrices.

THE COMPUTATION OF L MATRICES

For a given number of confidence degrees and a given scaling of the axis (place of the cut-off points), there is *one* L Matrix. Since it becomes more and more difficult to compute the L matrices as the number of confidence degrees increase, a computer FORTRAN program has been written to undertake the work.

How the FORTRAN Program Works

The program fixes TI0 and TC0 equal to 0, and TI1 equal to -1. It computes the resulting TC1. If the resulting TC1 is an integer, those values (TI1 and TC1) are accepted and the same procedure is followed for TI2 and TC2, with TI2 being fixed equal to -2. If the resulting TC1 is not an integer, TI1 is fixed equal to -2 and so on, until TC1 becomes an integer. As can be seen, a great number of steps could be necessary to find all the values for a given matrix. Fortunately, the computer does not get bored by iterative work.

Another problem is the definition of "integer". More exactly, should we accept values like 3.99 as integers and round them up to avoid big numbers? Each user has to answer this by fixing the threshold for rounding:
±.01 ? ±.05, etc.

How to Introduce the Data

The data needed by the L matrix FORTRAN program must be presented in the following order:

- The number of degrees of confidence (including level 0, if any), maximum = 100.
- The maximum number of iterations allowed (default option = 200).
- The threshold for rounding (default option = .05).
- The upper limits of the areas on the probability axis, expressed in percents (e.g., 25.0). Since 0.0 is always the lowest limit, it must not appear among the data. The last upper limit is always 100. The number of upper limits is the same as the number of degrees of confidence.

The FORTRAN Program

Figure 2.13 represents a FORTRAN program to compute L matrices. The user has to provide at least two cards for each matrix. The first card will contain three parameters and the second card will contain the upper limits of the confidence zones on the probability axis (see example in output in Figure 2.14).

```
C---------------------L MATRIX---------------------
C--------------------D LECLERCQ---------------------
C
      DIMENSION P(100),A(100),B(100),TC(100),TI(100),TRY(100)
C     A AND B WILL BE USED IN THE EQUATION (Y=AX+B) OF EACH SEU FUNCTION
C      SEU  = 'SUBJECTIVE EXPECTED UTILITY'
C
C----------- INPUT
C
    1 READ(5,500,END=3)NDC,NITER,TRESH
  500 FORMAT(I3,I4,F5.4)
      WRITE(6,603)NDC
  603 FORMAT(///'0NUMBER OF DEGRES OF CONFIDENCE = ',I3)
      WRITE(6,607)NITER
  607 FORMAT('0NUMBER OF ALLOWED ITERATIONS = ',I5)
      WRITE(6,608)TRESH
  608 FORMAT('0TRESHOLD FOR ROUNDING TO INTEGER = ',F6.4)
C
      READ(5,501)(P(I),I=1,NDC)
  501 FORMAT(20F5.1)
      WRITE(6,604)
  604 FORMAT('0LIMITS ON THE PROBABILITY AXIS :    0.0')
      DO 6 I=1,NDC
    6 WRITE(6,605)P(I)
  605 FORMAT(' ',34X,F5.1)
C---------- PROCESSING
      WRITE(6,609)
  609 FORMAT('0CONFID. *     TARIFFS     *    EQUATION'/' DEGREE  *'2X',
     1'TI(I)    TC(I) *  A(I)    B(I)'/1H ,44('-'),'*')
      B(1)=0
      TRY(1)=1
      A(1)=(TRY(1)/P(1))
      TC(1)=(100*A(1))-TRY(1)
      I=1
      TI(1)=-TRY(1)
      WRITE(6,601)I,T (1),TC(1),A(1),B(1)
  601 FORMAT(' ',I3,'     *',F6.1,F10.4,' *',F8.4,F10.4,' *')
      NDC1=NDC-1
      DO 2 I=2,NDC1
      NN=0
      K=I-1
C---------- TEMPTATIVE VALUE(TRY) FOR TI(I)
      TRY(I)=TRY(K)+1
      B(I)=(P(I)*A(K))-TRY(K)
    4 CONTINUE
      A(I)=(B(I)+TRY(I))/P(I)
      TC(I)=(100*A(I))-TRY(I)
      TI(I)=-TRY(I)
      WRITE(6,601)I,TI(I),TC(I),A(I),B(I)
C     TEST OF TRESHOLD FOR ROUNDING
      NN=NN+1
      W3=TC(I)+TRESH
      IW3=W3
      W4=IW3
      DIF=ABS(W4-W3)
      IF(DIF.LT.TRESH)GO TO 2
      TRY(K)=TRY(K)+1
      IF(NN.GT.NITER)GO TO 1
      TRY(I)=TRY(K)+1
      GO TO 4
    2 CONTINUE
C-----------OUTPUT
      WRITE(6,606)
  606 FORMAT(' ',46('-'),'*')
      WRITE(6,602)
  602 FORMAT('0',///' DEFINITIVE MATRIX')
      N=NDC-1
      DO 5 I=1,N
    5 WRITE(6,600)I,TI(I),TC(I)
  600 FORMAT(' ',I3,F6.1,F10.4)
      GO TO 1
    3 STOP
      END
```

Fig. 2.13. FORTRAN program to compute L tariff matrices.

The Output of the Program

```
NUMBER OF DEGRES OF CONFIDENCE =   4

NUMBER OF ALLOWED ITERATIONS =   200

TRESHOLD FOR ROUNDING TO INTEGER = 0.0100

LIMITS ON THE PROBABILITY AXIS :    0.0
                                   25.0
                                   50.0
                                   75.0
                                  100.0


CONFID. *     TARIFFS     *    EQUATION
DEGREE  *   TI(I)    TC(I) *     A(I)     B(I)
-------------------------------------------------
   1    *  -1.0     3.0000 *  0.0400     0.0    *
   2    *  -2.0     4.0000 *  0.0600     1.0000 *
   3    *  -3.0     4.3333 *  0.0733     2.5000 *
   3    *  -4.0     4.6667 *  0.0867     2.5000 *
   3    *  -5.0     5.0000 *  0.1000     2.5000 *
-------------------------------------------------


DEFINITIVE MATRIX
  1  -1.0     3.0000
  2  -2.0     4.0000
  3  -5.0     5.0000
```

Fig. 2.14. Typical output of the FORTRAN program presented in Fig. 2.13.

Numerous examples of L Matrices are given at the end of this chapter.

CONTINUOUS FORMULAE FOR COMPUTING D TARIFFS

Instead of using areas on the probabilty axis, one can refer to the continuum and express the tariffs by a function of p (the probability of success).

Basic Formulae

Van Naerssen (1965) suggested continuous formulae, referring to them as two *quadratic* solutions:

$$TC = A - B_q^2 \qquad \text{and} \qquad TI = A - B_p^2$$

where A and B are arbitrary constants, and $q = 1 - p$.

For instance, if A = 10.125 and B = 12.5, for p = .10 and q = .90,

TC = 10.125 - (12.5 . 0.81) = 10.125 - 10.125 = 0

TI = 10.125 - (12.5 . 0.01) = 10.125 - 0.125 = 10.

Table 2.8 presents TC and TI computed in the same way for p equal to .30, .50, .70 and .90.

TABLE 2.8. TC and TI Values of Van Naerssen's Quadratic Solutions for A = 10.125 and B = 12.5.

p	0-20	20-40	40-60	60-80	80-100
TC	0	4	7	9	10
TI	10	9	7	4	0

Figure 2.15 presents the graphic representation of the five SESQ functions:

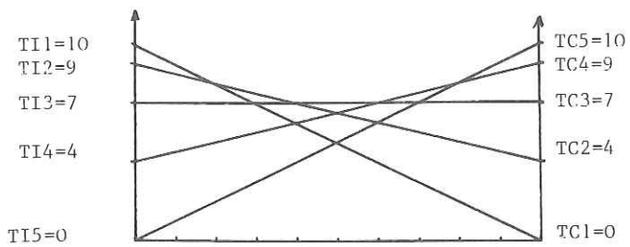


Fig. 2.15. Five SESQ functions for Van Naerssen's Quadratic Solutions, for A = 10.125 and B = 12.5.

In the following example (see Table 2.9 and Figure 2.16), TC and TI values have been computed for ten values of p : 0.05, 0.15, 0.25, etc.:

TABLE 2.9. TC and TI Values of Van Naerssen's Quadratic Solutions for A = 45.125 and B = 50.

p	.05	.15	.25	.35	.45	.55	.65	.75	.85	.95
TC	0	9	17	24	30	35	39	42	44	45
TI	45	44	42	39	35	30	24	17	9	0

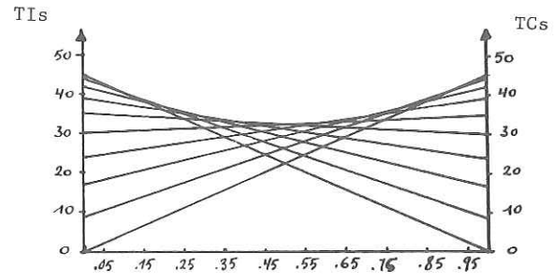


Fig. 2.16. Graphic representation of the ten SESQ functions defined (in TABLE 2.9.) for ten p values.

Note that all the values are positive. To obtain negative values, it is sufficient to change the A value. Table 2.10 presents such a t matrix.

TABLE 2.10. TC and TI Values of Van Naerssen's Quadratic Solutions for A = 3.125 and B = 12.5

p	0-20	20-40	40-60	60-80	80-100
TC	-7	-3	0	2	3
TI	3	3	0	-3	-7

Whereas these are D tariffs (that is, computed according to decision theory), they are not E tariffs (Educational ones). Remember that in Educational tariffs, all the TC values are positive, all the TI values negative, and the TO (tariff of omission) value is null.

An Educational Formula

The above tariffs are inadequate for educational purposes either because some TI values are positive or because some TC values are negative. With the five zones presented in Table 2.7, let us keep the TCs unchanged (0, 4, 7, 9, 10) and let us lower all the TIs by, say, 10 points. The TC and TI values obtained are presented in Table 2.11.

The resulting graph appears in Figure 2.17.

TABLE 2.11.

p	0-20	20-40	40-60	60-80	80-100
TC	0	4	7	9	10
TI	0	-1	-3	-6	-10

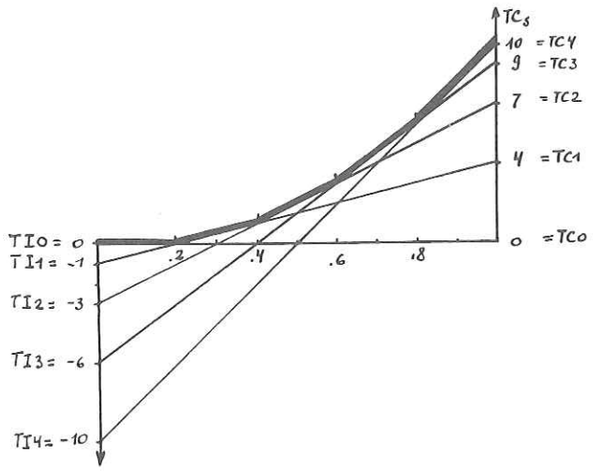


Fig. 2.17. Graphic representation of five SESQ functions for the TI and TC values presented in TABLE 2.11.

The formula is unchanged for TC: $TC = A - B\,(1-p)^2$

It is slightly changed for TI: $TI = A - B_p^2 - 10$

General formulae are:

$$TC = A - B\,(1-p)^2 + C$$
$$TI = A - B_p^2 + D$$

where C and D are arbitrary constants (in Figure 2.17, C = 0 and D = -10).

The darkened line in Figure 2.17 is composed of the portions of SESQs that exceed the others. This line is referred to as the *maximization line*.

If the TCs and the TIs are plotted at the vertical of the middle of their zone, we obtain the diagram of Figure 2.18.

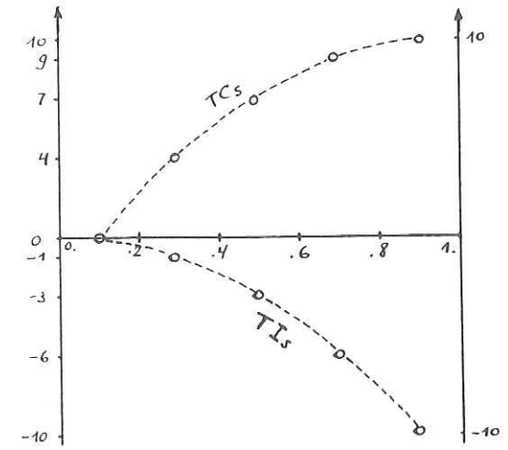


Fig. 2.18. Plotting of TC and TI values for five values of p (see TABLE 2.10.).

Parameters of Educational Formulae

In the above example, TC0 = TI0 = 0. Of course, it is possible to conceive of tariffs where TC0 > TI0. What will consequently change is the Relative Importance of Confidence (R.I.C.) compared with correctness or incorrectness of the answer.

In the four following examples (see Figure 2.19), C and D values have been changed in the general formulae, whereas A remains equal to 10.125 and B equal to 12.5.

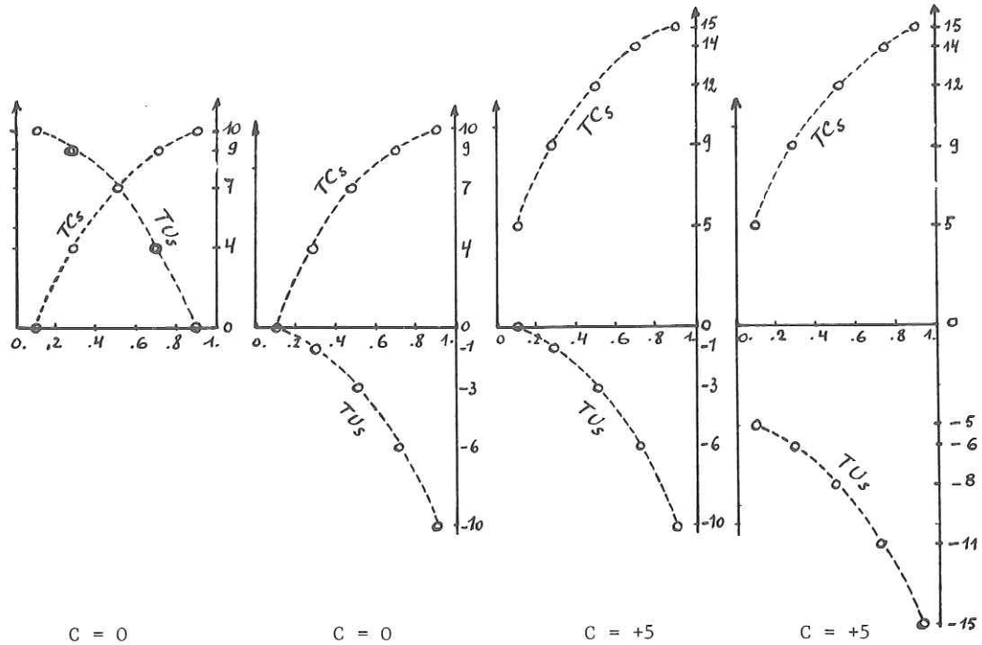
It is possible, too, to vary the formula itself, in order that, for instance, TC becomes a linear function of p: $TC = Ap + C$.

In the L matrix presented in Table 2.7 (were TC1 = 3, TC2 = 4, and TC5 = 5), the corresponding p values are respectively 0.375, -0.625, and 0.875. For example, TC1 = (4 x 0.375) + 1.5 = 1.5 + 1.5 = 3.

The values of TI have to be computed by a square function.

It must be noted that, since the tariffs of the first zone (from 0 to 0.25) have been fixed at 0 (TC0 = TI0 = 0), the previous formula can only be applied to the zone of the axis ranging from 0.25 to 1. That is the reason why a vertical axis is drawn at the 0.25 limit. The picture formed by this vertical limit, the curve of TCs and the curve of TIs have the general form of the letter K. Hardy (1980-81) has provided a penetrating and

comprehensive presentation of K scales. He stresses the point that the choice of A, B, C and D parameters depends upon the choice of the general shape of the maximization line (or curve).



C = 0 D = 0 | C = 0 D = −10 | C = +5 D = −10 | C = +5 D = −15

Fig. 2.19. Plotting of TC and TI values for A = 10.125 and B = 12.5.

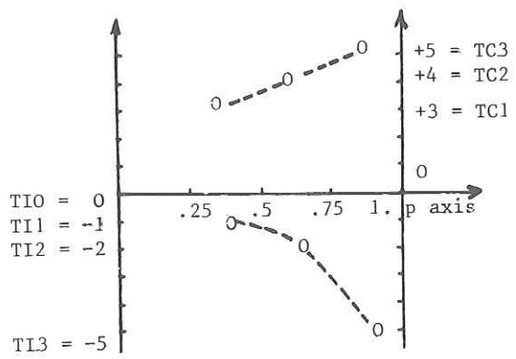


Fig. 2.20. Plotting of TC and TI values appearing in TABLE 2.7.

THE PROBLEM OF COMPLEX FORMULAE

Rippey (1970) administered a test to 1000 high school students. His instructions were as follows:

> Each of the questions or incomplete statements in this test is followed by suggested answers. Assign a number from 0 to 9 to each suggested answer depending on how strongly you feel the answer is correct. If you believe that only one suggested answer is correct, mark that answer with a 9 and mark the other(s) with zeros. If you like the suggested answers equally, assign the same number to each.

He then showed various examples of answers and ended his instructions by:

> Your paper will be scored in such a way that you will get a higher score by estimating your degree of confidence and reporting it accurately. Guessing in any form will lower your score. If you are uninformed about the question and have no preference for the suggested answer, you will obtain your highest score by honestly distributing your confidence across all the options...

Note that this last sentence is aimed at completely ignorant students and that the instructions indicate to these students how to answer where, in normal circumstances, they would have omitted.

The crucial point is that Rippey does not reveal to the students exactly what the tariffs will be and how they are computed. In fact, he uses five distinct formulae to obtain five different scores to each question (SQ) for each student.

In these formulae:

pc is the probability attributed by the student to the correct alternative.
pi is the probability attributed by the student to alternative i.
ri is a reference probability attributed to alternative i by a group of experts.
k is the number of alternatives in the question.

Formula 1 (the simplest) : $SQ = pc$

Formula 2 (logarithmic) : $SQ = (2 + 10g_{10}pc)/2$ (except when $pc < .01$; then $SQ = 0$).

Formula 3 (spheric) : $SQ = pc \,/\, (\sum_{i=1}^{k} (pi)^2)^{1/2}$

Formula 4 (euclidian) : $SQ = 1 - [\sum_{i=1}^{k} (pi-ri)^2)^{1/2}] \,/\, \sqrt{2}$

Formula 5 (inferred choice) : SQ = 1 if pc > pi for any alternative.

He concludes that, in a situation of ignorance about the scoring formula, students attribute their confidence indexes *on the basis of the simplest formula*.

But it should be noted that the fifth formula (in fact, the ordinary way of scoring without confidence indexes) has the worst reliability index.

Rippey's conclusion is that *most scores computed using esoteric scoring functions will have an error component which is due to the subject's lack of understanding of the scoring system*.

He illustrates this error component by the shaded surface in the following graphs (Figures 2.21 and 2.22). The abcissa represents pc and the ordinate SQ. If the student believes that SQ will be equal to pc, and if his belief is true, his observations will take place on the diagonal. But, if the SQ is computed another way, there is a discrepancy that will increase the error component of SQ.

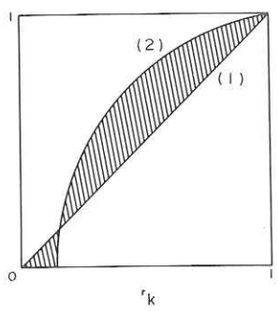


Fig. 2.21. Graph of Scores for Functions 1 and 2.

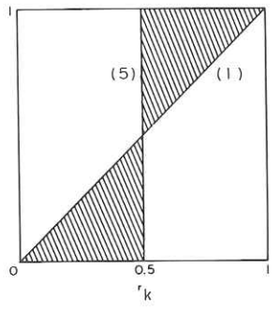


Fig. 2.22. Graph of Scores for Functions 1 and 5.

He concludes as follows:

> *I would be most inclined to score all tests having unique correct responses using Function 1*, the probability assigned to the correct answer, for the following reasons: (a) reliability is higher than that obtained by inferred choice methods, (b) no expense or scoring complications are involved as is the case with the other functions, and (c) the idea of having a student state his degrees of belief rather than asking him to be dogmatic about his uncertain preferences seems to be more realistic and more honest for both the student and the instructor.

Teachers should be aware of the transparency of a scoring system. The formula itself is often too complicated. Showing the whole scale of tariffs to the students is the most direct solution.

THE RELATIVE WEIGHT OF CONFIDENCE INDEXES IN TEST SCORES

In some tariff scales, the difference between the least TC and the least TI is important. In other tariff scales, TCi is very different from TCj, whereas there is not very much difference (the C and D values in above formulae) between the least TC and the least TIl.

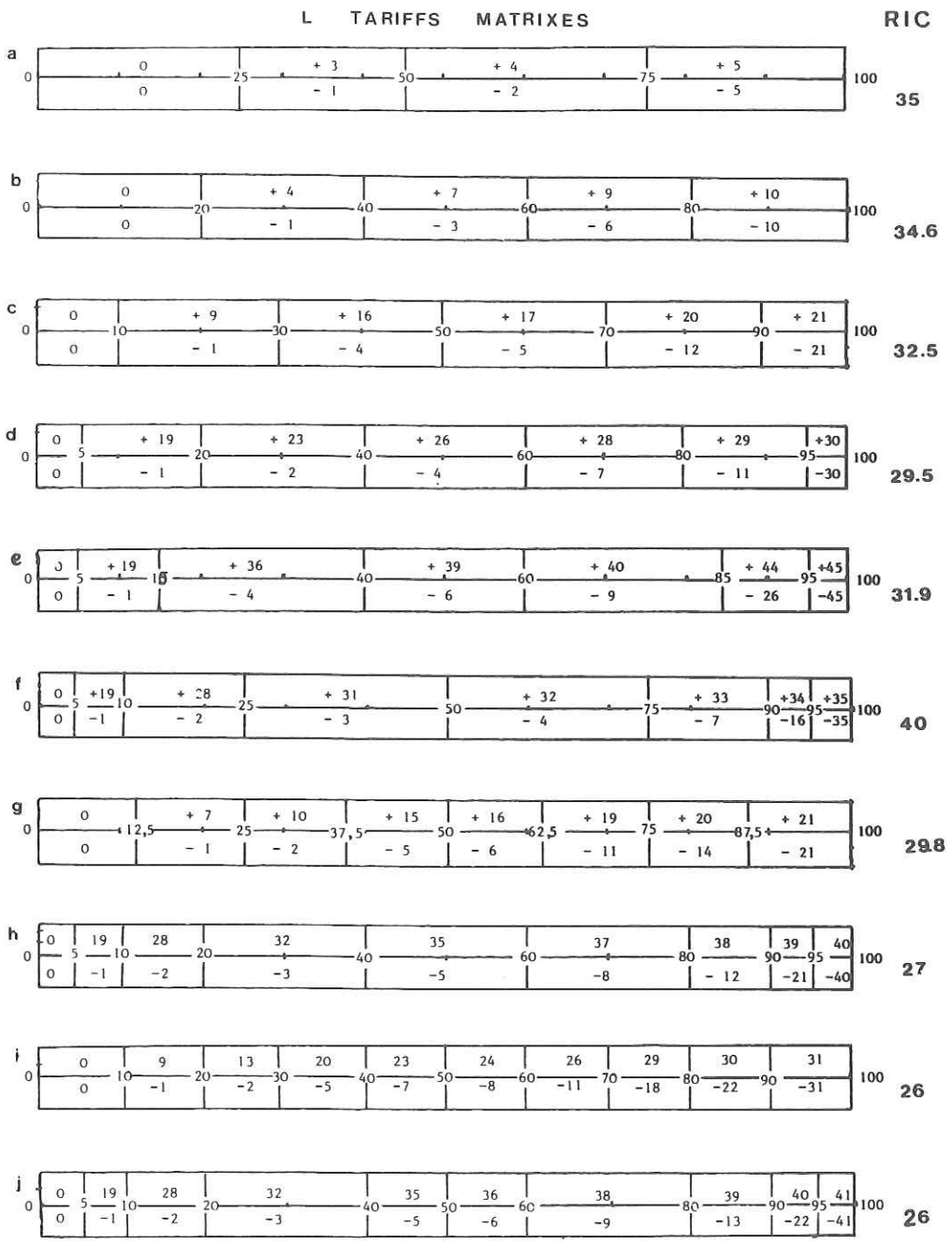
The three scales presented in Table 2.12 illustrate this phenomenon.

TABLE 2.12. Tariff Scales varying from each other according to the Difference between TC1 and TI1 (great difference in scale A, low difference in scale C, intermediate difference in scale B) and according to the difference between TC3 and TC1 (great difference in scale C, low difference in scale A, intermediate difference in scale B).

	SCALE A			SCALE B			SCALE C	
	TC	TI		TC	TI		TC	TI
C0	0	0	C0	0	0	C0	0	0
C1	+1,9	-1,9	C1	+1	-1	C1	+0,1	-0,1
C2	+2	-2	C2	+2	-2	C2	+2	-2
C3	+2,1	-2,1	C3	+3	-3	C3	+3,9	-3,9

A simple analysis of variance reveals what proportion of variance is due to confidence indexes (this variance can only appear in the interaction term) compared with the total variance. The ratio between the two sums of squares is considered as the relative importance of confidence indexes (RIC). The values of RIC appear in Table 2.13 at the right hand side of ten scales with their L matrix of tariffs.

TABLE 2.13. TC and TI Values of ten L Educational Tariff Matrices for Arbitrary fixed Zones of the Probability Axis.
The RIC Value is the relative importance of confidence in the scoring matrix (RIC is expressed in percentages).



CHAPTER 3

THE VALIDITY OF CONFIDENCE MARKING PROCEDURES

THE VALIDITY PROBLEM

The problem of the validity of confidence indexes is often poorly stated because the score for a question (SQ) and the score for a test (ST), computed with special confidence tariffs are *not measures* of the student's ability.

Let us stress this point: the number of correct answers (that is the S t scale) is a *measure* of the student's ability and the score computed with the G t scale (with a correction for guessing) is another *measure* of this ability. It makes sense to apply classical validity coefficients (correlations, etc.) to try to find out which one (SST or SGT) is the best measure.

Conversely, SCT, that is, the score on a test computed with a C t scale is *not a measure* but a mixture of two different measures: the measure of the student's ability in the content and the measure of the student's ability in self estimation. We shall see that there are various indexes of this ability: PSY (or) indexes (of realism, of calibration, etc.).

The SCT score is the result of *an aggregation* of those two measures, with the relative importance (RIC) depending on the tariffs matrix. This SCT is *a payoff*, a reinforcement, not a measure.

It is therefore pointless to try to find out whether this new score (SCT) is a better measure than other classical scores. A lot of (vain) research studies have been undertaken on this point and produced contradictory results. Some researchers observed an increase of validity and a decrease of reliability whereas other researchers observed the contrary. These conflicting and apparently nonsensible results can be explained. If the student is a good estimator of his own capacity, and if he tells the truth, then we obtain, from his confidence indexes, more information about his ability in the domain. Of course, if he is a bad estimator, or if he biases what he thinks when he expresses his confidence, we get confusing information. The problem here is to distinguish trustworthy information from non-trustworthy information.

This chapter will focus on the questions "Do students bias their estimation when they express it? What can teachers and experimenters do to cope with this problem?" These are two crucial questions about the validity of

confidence marking. Coming back to the general model (see Figures 1.5 to 1.9), we could state the problem in following terms: "Are the osp_s an unbiased expression of the esp_s ? If not, why not?"

Decision theorists (mostly economists and psychologists) have described a lot of strategies of behaviors people adopt. Telling the truth is only one of them. The test instructions and tariff matrix can be conceived in such a way that telling the truth is objectively the most interesting strategy (from the payoff point of view) because it maximizes the expected score. Even in this case, oral or written explanations given by the teacher are frequently useless. Students are not convinced of the superiority of the strategy "tell the truth" *unless they have experienced it*.

Below, an experiment is related where favorable circumstances reveal clearly the role of contingencies of reinforcement on the expression of confidence in test situations.

A REVEALING EXPERIMENT

Institutional Background

The two experiments reported here took place at the Belgian Air Force (BAF) Technical School located at Saffraanberg, in 1971 and in 1972. At this school, an item bank has existed since 1970 and currently contains more than 30,000 items in about thirty content areas, and several hundred tests are administered each year. We were deeply involved in creating and developing the item bank (Leclercq, 1973, 1975).

Multiple choice items produced by teachers are selected, checked, typed, coded and stored by a special team. Any teacher can obtain a test made of a desired subset of items in the bank. This test is then reproduced in the required number of copies by the technical team. The students answer the test with a pencil on a preprinted grid (see Figure 3.1).

Answers																				
Questions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Confidence degrees																				

Fig. 3.1. Typical preprinted grid used as answer sheet for students.

A second strip of (self copying) paper is joined to the original. At the end of the test the students give the copy to the teacher and keep the original either to correct collectively or as evidence in case of a complaint. The teacher sends the copies of the strips to the central team which punches the data and analyzes them using a program (EVAL) written in FORTRAN. The whole system has a great capacity and rapidity: several tests can be treated in a few hours, so that feedback to the students is rapid. Many teachers present several tests to the same group of students during a school year.

Such a situation is excellent for research purposes since it enables the observer to deal longitudinally with the evolution of behaviors. We are grateful to P. Van Roy, who has run this center from the beginning, for this collaboration and his permission to use data for our research purposes.

The Experiment

In 1971, at the BAF Technical School, the test instructions presented an ordinal scale and an educational payoff matrix (but *not* a D matrix) shown in Table 3.1.

TABLE 3.1. Instructions and Tariff Matrix used in the 1971 Experiments.

Give confidence degree	Number	TC	TI
If you are not sure at all.	0	0	0
If you are weakly sure.	1	+1	-1
If you are moderately sure.	2	+2	-2
If you are strongly sure.	3	+3	-3

At that time, we were not yet aware of either the importance of a probabilistic (and not ordinal) definition of confidence degrees or of the necessity to use D matrices.

Groups of students received several tests during the school year. We shall focus on the results of 62 students from four classes that received 14 tests in mechanics, each test having about 25 questions, in the period from September 1971 to June 1972. Only 53 students from the 62 finished the whole year.

The Method of Analysis

The focus analysis is in the pattern of uses (U pattern) of the different confidence degrees by a given student for a given test. Such a pattern is defined, here, by four numbers:

the number of uses of confidence degree 1 (NUC1)
the number of uses of confidence degree 2 (NUC2)
the number of uses of confidence degree 3 (NUC3)
the number of uses of confidence degree 0 (NUC0 or ommission).

For each student, NUC0, NUC1, NUC2, and NUC3 constitute his U pattern for one test.

The analysis was not undertaken on 742 U patterns (53 students times 14 tests), but only on 703 U patterns because of 39 absences. A typical individual pattern is the one where only the greatest confidence degree has been used. Such a U pattern is compatible with the MAXIMAX strategy (see below), which implies that this U pattern could have been elicited by the use of this irrelevant strategy. Note that *compatibility does not mean causality*.

In order to be able to analyze the data of the experiment, it is necessary to consider the classical strategies described either by economists or by psychologists (both are discussed below).

INAPPROPRIATE STRATEGIES DESCRIBED BY ECONOMISTS

All the strategies presented below used the payoff matrix presented in Table 3.2.

TABLE 3.2. Typical Tariff Matrix (has been used in the 1971 experiment).

	0	25	50	75	100
TCs	0	+1	+2	+3	
TIs	0	-1	-2	-3	

Two Extreme Criteria

The MAXIMAX criterion. This criterion is also called *maximum utility criterion* (Coombs, Dawes and Tversky, 1971, p. 141). It consists of choosing the possible act (here the confidence degree) that will produce the *best win* (here the best tariff) in the case of success. Undoubtedly, this is an optimistic criterion and, in our example, would lead to choosing the greatest degree of confidence (number 3).

The MAXIMIN criterion. This criterion is also called *maximin utility criterion* (Luce & Raiffa, 1966, pp. 23-31) or even the Wald criterion. It consists of choosing the possible act (here the confidence degree) that will produce the *least loss* (here the best tariff) in the case of failure. Clearly this is a pessimistic criterion that, in our example, would lead to choosing the lowest degree of confidence (number 1, or number 0 when 0 is not an omission). It is often said that this criterion warrants *the best "worst state"*.

Comments on these two criteria. Coombs, Dawes, and Tversky (1971, p. 141) note that "because things, however, are usually neither as bad as we feared nor as good as we hoped, it might be advisable to weigh the best and the worst". This recommendation is implemented in the following three criteria.

The Intermediate Criteria

The IGNORANCE criterion. This criterion, sometimes attributed to Bernouilli (1654-1705) is often called *the principle of insufficient reason, or the principal of equiprobability*, or even Laplace criterion.

In this criterion, each state of nature (here success and failure) has an equal probability of occurring. Consequently, the preferred act (here confidence degree) is the one that gives the maximal sum of consequences (here the maximum TC + TI), or the maximal mean (sum/2) of these two tariffs.

In our example, all the degrees are equivalent (their TC and TI sum to 0).

The PESSIMISM/OPTIMISM criterion. This criterion, due to Hurwicz, assumes that the student attributes to each state of nature (here success or failure) a given probability. This value is the same for all the questions and depends only on the subject (it can be considered as a personality trait). So, pessimistic persons will attribute a (permanent) weak probabiltiy to success, whereas optimistic persons will fix this probabilty to a (constant) high level.

It may be noted that if the a priori probability of success is 1, this criterion confounds itself with maximax criterion; conversely, when the a priori probability of success is 0, it is confounded with the maximin criterion.

In our example, a given student would always use the same confidence degree, regardless of the question.

The MAXIMAX REGRET criterion. This criterion presented by Savage (1951), is an over-refinement of the *maximin criterion*. It is pessimistic since it is focused on unfavorable events. From the original payoff matrix, Savage builds another matrix called the "*regrets matrix*" where each cell contains the deviation from the gain that would be given by the best decision (if the state of nature were known).

In our example, this best gain is 3 in case of success and 0 in case of failure. A regret matrix contains no positive value. In our example, these values are shown in Table 3.3.

TABLE 3.3. Regret Matrix for the Tariff Matrix presented in TABLE 3.2.

	TC	TI
Degree 0	<u>-3</u>	0
Degree 1	<u>-2</u>	-1
Degree 2	-1	<u>-2</u>
Degree 3	0	<u>-3</u>

For each possible act (here for each degree of confidence) the greatest regret (in absolute value) is underlined. The lowest of those underlined values decide which action to select (here degree 1 or degree 2).

Discussion About the Five Criteria

The five criteria given above are sometimes included in axiomatic systems (Chernoff, 1954 or Milnor, 1954), an overview of which can be found in Luce and Raiffa (1966, p. 297).

Coombs Dawes and Tversky (1971, p. 142) suggest a numerical example, in the case of five possible acts, where each action is the best one according to one of the five criteria.

TABLE 3.4. Particular Tariff Matrix where Each possible Act (A1 to A5) would be preferred according to a different criterion.

	Tariffs in case of :			
Possible acts	States of nature : 1	2	3	Criterion justifying the choice of each act :
A1	5	5	5	Maximim (WALD)
A2	10	0	0	Maximax
A3	9	2	2	p = 0,5 (HURWICZ)
A4	8	0	8	Ignorance (LAPLACE)
A5	6	1	4	Minimax regret (SAVAGE)

INAPPROPRIATE STRATEGIES DESCRIBED BY PSYCHOLOGISTS

Risk Taking and Need Achievement

For McClelland (1953, p. 79), need achievement (Need Ach) is "an affect in connection with evaluated performance". It can be the basis for evaluating the attractiveness of events where no tariffs are defined. In this context, Atkinson (1964) developed a theoretical model in which any risky situation presents two components: hope of success and fear of failure. His central hypothesis is that the more difficult the tasks, the greater the subject's satisfaction in the case of success. So, the *incentive value of sucess* (IVS) is equal to sq (that is 1 - sp). Moreover, each subject is characterized by a personal strength in terms of the need of success (NOS).

In a particular situation, the motivation (to choose one action), or the actualized need is called the *tendency to approach success* (TAS).

Atkinson's basic formula is:

$$TAS = NOS \times (sp \times IVS)$$

The tendency to approach success (TAS) will vary proportionally according to:

- the individual's need of success (NOS
- the (subjective) probability of success (sp)
- incentive value of success (IVS), that is, the (subjective) probability of failure (sq).

Since sp.IVS is maximal for average difficulty tasks (sp = sq = .50, so sp.IVS = .25), individuals with a great *Need Ach* will choose those average difficulty tasks. On the other hand, students who have a great motivation to avoid failure will choose easy tasks. This is, of course, a psychological version of the economic principle of maximizing subjectively expected utility (SEU). It resulted in a model of choice behavior called "probability preferences".

Probability Preferences

This model is not directly related to the situation of choosing a confidence degree where the student is not allowed to choose the difficulty of the items, and the choice between confidence degrees is not a choice between acts of various difficulties. For this reason, Atkinson's model has not been tested in the experiments that will be described below. Nevertheless, it could be tested when the teacher allows the students to choose among (weighted) items in order to reach a given total of weights.

Evidence of preferences of probabilities have been provided by numerous authors, e.g., Edwards (1953), Atkinson and Litwin (1960), Clark et al (1956), Smith (1963), Isaacson (1964) and Myers (1965). Other results show that in random games (where competency is useless), students who have a high "need ach" take low risks (Littig, 1954, Hancok & Teevan, 1964; Raynor & Smith, 1965).

Risk Variance Preferences

Coombs (1967) developed a psychological theory of choices where each student had his ideal point of risk. If the risk of a confidence degree is defined as TC - TI for a particular degree, we see that, in our classical matrices of tariffs, lower degrees of confidence imply low risk, whereas higher degrees imply high risks. According to Coombs, a given student's ideal point on the probabilty axis depends upon its distance from this (unknown) ideal point. In order to test Coombs' theory, we computed the patterns resulting from the set of all the possible ideal points. We then detected those "compatible" patterns amongst the observed ones. Note that other authors (i.e., Slovic & Lichtenstein, 1968) have suggested other approaches to risk variance preferences.

THE RESULTS OF A LONGITUDINAL EXPERIMENT

Let us recall the crucial question "Do students use strategies (listed earlier) other than telling the truth?" If the answer to this question is yes, confidence responses are not trustworthy and their usefulness has to be questioned. In the following reasonings, the "suspect patterns" will be considered as "bad" patterns (i.e., possible undesirable strategies).

Patterns Compatible with Optimistic Strategy (MAXIMAX)

During the 14 successive tests administered in 1971 at the BAF Technical School, only 10 students out of 53 exhibited a pattern with only one confidence degree: the highest. Seven out of ten of them used it only during one test. Nevertheless, it can be considered that other U patterns are compatible with the MAXIMAX criterion; for instance, the U pattern where NUC2 + NUC3 = NQ, (using only the two highest degrees), as well as the U pattern where NUC0 = 0 (no use of omission), and as the U pattern where (NUC3/NQ) > .95 (very important use of the higest degree).

Figure 3.2 presents the evolution, over the fourteen tests, of those U patterns compatible with the MAXIMAX criterion.
The progressive decrease of U patterns compatible with undesirable strategies indicates that the problem of validity (credibility of subjective probabilities) must be stated in terms of repeated testing for trained students.

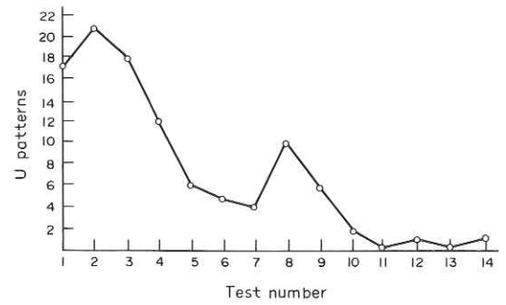


Fig. 3.2. Evolution over 14 successive tests of the number of observed U patterns compatible with MAXIMAX criterion (in the 1971 experiment where 53 patterns were available for each test).

Patterns Compatible with Pessimistic Strategies

Strategies 2 (MAXIMIN) and 3 (IGNORANCE), even when broadly interpreted, seem to have elicited very few U patterns, if any. A few U patterns compatible with strategy 5 (MAXIMAX REGRET) appeared only during the four first tests (see Figure 3.3).

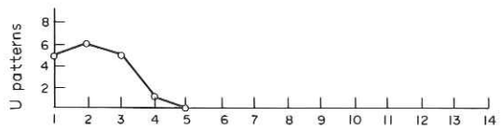


Fig.3.3. Evolution over 14 successive tests of the number of observed U patterns compatible with MINIMAX REGRET criterion (on 53 patterns, in the 1971 experiment).

Patterns Compatible with Risk Variance Preferences

Turning to psychological theory, we noted U patterns compatible with Coombs' unfolding theory, that is a strategy where some risk variances are systematically preferred. Again, the number of those patterns decreased with repeated testing (see Figure 3.4).

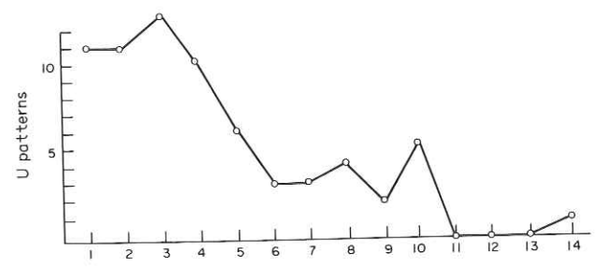


Fig. 3.4. Evolution over 14 successive tests of the number of observed U patterns compatible with Coombs' Preference of Risk Variance Criterion (described by unfolding theory), in the 1971 experiment (on 53 patterns at each test).

An Explanation of the Evolution

The same tendency (progressive decrease) has been observed for all the "undesirable" U patterns, whereas the teachers (unaware of this problem) had no specific action on this point.

The question then arises "Why did the students desert those strategies?" The most plausible explanation whould seem to be operant conditioning: students' behaviors were controlled (we could say regulated) by their consequences (the payoffs or reinforcements). It can easily be proved (see "decision theory" in the previous chapter) that undesirable strategies were less effective (pay less) than the (only) strategy based on decision theory (D strategy), that is maximizing one's subjectively expected score to the question (SESQ).

The problem now is: "Did the number of D strategies increase progressively with repetitions of tests?" This, in turn, gives rise to a new problem: Are D strategies observable through U patterns? At this point, we have been very lucky and serendipity (finding something you were not searching for) helped a lot. Indeed, when D matrix is applied, there is no typical U pattern to represent a D strategy and any U pattern could have been caused by (or is compatible with) a D strategy. Fortunately, we did not use a D matrix.

A Confirmation of the Explanation

In 1971, we were not aware of the importance of using a D matrix and the tariff matrix used (+1, +2, +3, -1, -2, -3) *was not a D matrix*; it was only an Educational one! Figure 3.5 presents the four SESQ functions, one for each degree of confidence (0, 1, 2 and 3). It appears that two of the functions are never optimal: SESQ1 and SESQ2.

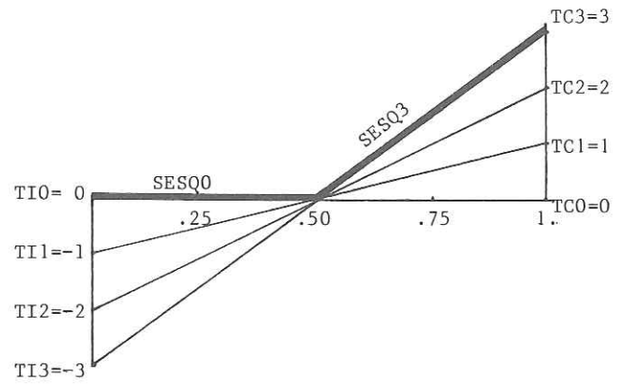


Fig. 3.5. Graphic representation of the four SESQ functions for the tariff matrix presented in TABLE 3.2. used in the 1971 experiment.

Since such a graph had not been seen by either the experimenters, the teachers, or the students, the students could hardly have reached the (obvious) conclusion that confidence degrees 1 and 2 should not be used because the SESQ was higher if they used confidence degree 0 (when one's sp is lower than .50) and confidence degree 3 (when one's sp is greater than .50). This kind of conclusion is difficult to reach since when sp = .50 exactly, the four degrees of confidence are equally attractive.

At the time, not being aware of the arguments given above, we were puzzled when analyzing the scores of fourteen tests that *confidence degrees 1 and 2 were progressively abandoned*. In 1971-72, graphs appearing in Figure 3.6 show the big difference between the first three tests and the following ones.

During the next school year (1972-73), a D matrix of tariffs was used. The results, on 13 successive tests, showed the stability of the relative proportion of degrees of confidence 1 and 2 contrasted (see Figure 3.7) with their progressive decrease during the school year 1971-72.

A test facility decrease along the school year had an impact on the use of low degrees of confidence (0 and 1) since the correlation between test difficulties

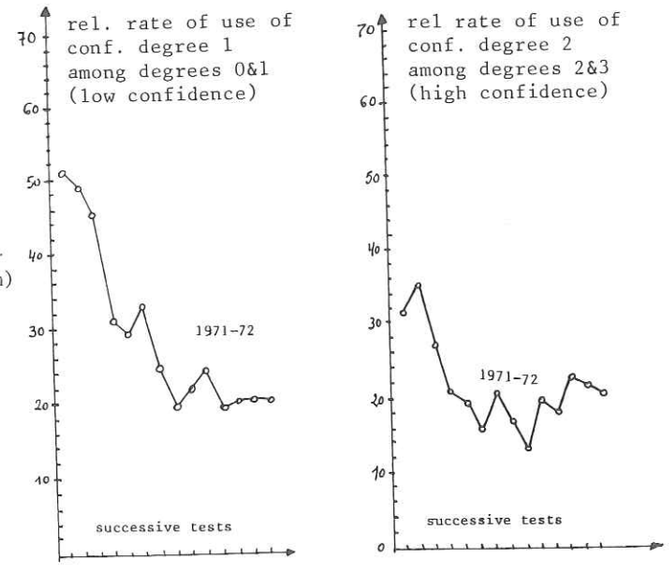


Fig. 3.6 Evolution through 14 successive tests of the relative rate of use of confidence degree 1 (left graph) and of confidence degree 2 (right graph) during the 1971-1972 experiment.

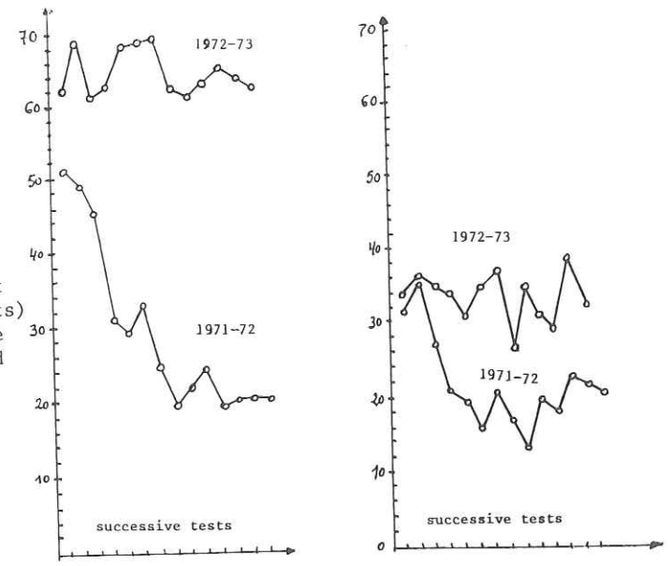


Fig. 3.7 Contrasts between the 1971-1972 experiment (using an E-tariff matrix) and the 1972-1973 experiment (using a D-tariff on 13 tests) on the relative rates of use of degree 1 (left graph) and degree 2 (right graph).

and the use of low degrees was .886.

Thus, the evolution of degree 1 among the low degrees (0 and 1) appeared to be a more valid indicator than the evolution of its absolute rate. The same can be said for the relative rate of the use of confidence degree 2 among the high degrees (2 and 3). It can be observed that, in the first tests, the relative rate of degree 0 and 1 were approximately equal (.50) and the relative rates were approximately 1/3 for degree 2 and 2/3 for degree 3. At the end of the series of tests, the relative rate of confidence degrees 1 and 2 stabilized around .20.

This is not only an average observation. Individual evolutions also confirm this tendency. This check was necessary since "an average curve rarely gives a correct image of any of the individual cases on which it relies" (Sidman, 1953).

Conclusion

All the data support the general hypothesis that during the first year an operant learning took place. This learning was largely unconscious, since students could hardly explain *why* they avoided degrees 1 and 2, though they could say vague things like "it pays more".

The observed curves are typical of operant avoidance. Figure 3.8 presents the average evolution of the relative rate of use of degree 1 (left hand side curve). The curve on the right is the number of shocks received by a white rat during successive periods of 15 minutes, in an avoidance conditioning situation.

The rat had to walk in a circular cage built for this purpose. When it interrupts the (photocell) ray of light, it postpones the shock of 15 seconds (see Beaujot, Didelez, Fontaine & Leclercq, 1966).
Comparing human and animal curves in the domain of behavior will appear a sacrilege to some people. However, we did it on purpose because we wanted to combat an old ambiguity related to operant conditioning. In physiological domains, everybody accepts the idea that some laws or phenomena are relevant for both humans and animals and nobody would be scandalized by a comparison of electrocardiograms from humans and animals.

"But the fundamental confusion," says Richelle (1970) as a reply to the French biologist, Chauvin (1967), "consists in seeing in conditioning a *category of behaviors* whereas it is a *mechanism* ... If the mechanism of conditioning is, in principle, extremely simple, it does not imply that the resulting behaviors are also simple."

The problem of the validity of derived score and measurement procedures will be treated in following chapters since it is a more theoritical problem that can now at last be discussed on a sound experimental basis.

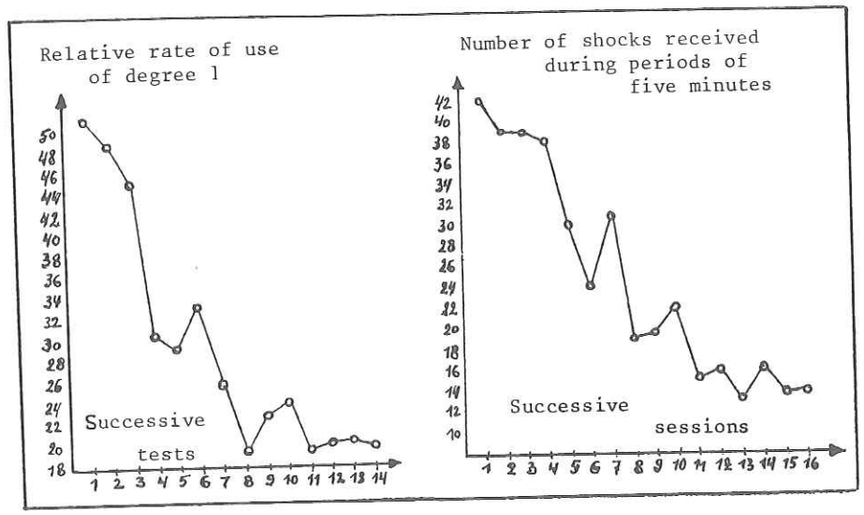


Fig. 3.8. Two typical curves obtained by operant conditionning (avoidance of behavior followed by unwanted consequences). The left hand side curve represents avoidance of a given confidence degree in human learning. The right hand side curve represents avoidance of electric shocks by a rat.

CHAPTER 4

THE USE OF CONFIDENCE MARKING TO EVALUATE STUDENTS

The quality of the student's self estimation is a central concern in confidence marking (see Figure 1.9). Four types of indexes will be proposed: indexes of coherence, of realism, of calibration, and of efficiency. These indexes will be referred to as ψ (PSY) indexes. Procedures will be suggested for rapid computation of some of them.

The influence of personality on self assessment has been stressed for a long time (Wiley & Trimble, 1936; Hevner, 1932; Swineford, 1941; or, more recently, Jacobs, 1971).

It is obvious that this kind of information on the individual student can be used formatively by the teachers. New test scores based on confidence marking need careful interpretation, and basic concepts of measurement theory are needed.

TABLE 4.1. Instructions that will be used as Example to develop the Principles of Chapter 4. Those instructions were already presented in Fig. 2.8. and in TABLE 2.7.

In the following examples, the instructions in effect are the following ones (see TABLE 4.1.).

- Four degrees of confidence are available : 0, 1, 2 and 3.
- The TIs are, respectively, 0, -1, -2 and -5.
- The TCs are, respectively, 0, +3, +4 and +5.
- The cutpoints on the probability axis are .25, .50 and .75.

THE MEASUREMENT OF COHERENCE

In Chapter 3, a pattern of use (NUC0, NUC1, NUC2, NUC3) was examined for each student's answers to a given test. Let us recall that NUC1 means "number of uses of confidence degree number one".

The coherence index (ψco or PSYCO) is based on another pattern, that of the *rates of success* (RS) for each confidence degree (i.e., RSC1, RSC2, etc.). This rate of success is computed by the simple formula: $RSC_i = \frac{NSC_i}{NUC_i}$

where NSC_i is the number of successes among responses given with confidence degree i.

Suppose that a student was given a test containing 20 items and used the confidence degrees as in Table 4.2.

TABLE 4.2. Example of Uses of Confidence Degrees 1, 2 and 3 in a Test containing 20 Items.

The student used :

- Confidence degree 1 for 6 responses (NUC1 = 6), two of them being correct (RSC1 = 2/6 = .33).
- Confidence degree 2 for 4 responses (NUC2 = 4), three of them being correct (RSC2 = 3/4 = .75).
- Confidence degree 3 for 10 responses (NUC3 = 10), nine of them being correct (RSC3 = 9/10 = .9).
- Confidence degree 0 was never used.

In the example of Table 4.2, it can be observed that .33 < .75 < .9, that is:

$$RSC1 < RSC2 < RSC3$$

Inequality is a sign of *strong coherence*. Cases where only RSC1 < RSC3 will be called weak coherence. Cases where RSC1 > RSC3 will be called incoherence. So, ψco (or PSYCO) has only three nominal values: strong coherence, weak coherence, and incoherence. PSYCO is a very weak index for at least two reasons.

First, it happens frequently that a test has only a few items (NQ < 30). Each NUC_i is, consequently, very low and the RSC_i must have a large standard error of measurement.

Second, this PSYCO index lacks accuracy. For instance, in all the examples of Table 4.3, the PSYCO value is "strong coherence", but, obviously, all of these situations are *not* equivalent.

TABLE 4.3. Five Contrasted Situations of Strong Coherence (Strict Order in the three Rates of Success of Confidence Degrees 1, 2 and 3).

	RSC1		RSC2		RSC3
Situation 1	.37	<	.47	<	.68
Situation 2	.28	<	.62	<	.95
Situation 3	.65	<	.75	<	.87
Situation 4	.70	<	.72	<	.75
Situation 5	.37	<	.62	<	.87

THE MEASUREMENT OF CALIBRATION

In the given instructions, the confidence degree 1 covers probabilities ranging from .25 to .50. The central value (CV) of this confidence zone is .375. The central values corresponding to the instructions of Table 4.1 appear in Table 4.4

TABLE 4.4 Central Values of Confidence Degrees 0, 1, 2, and 3.

CVC0 = .125
CVC1 = .375
CVC2 = .625
CVC3 = .875

The basic principle of the measurement of calibration is the comparison between the (observed) RSCs and the (theoretical) CVCs, that is, the Mean Error of Estimation (MEE) for each degree.

In the example of Table 4.2, the MEE values are easily computed (see Table 4.5).

TABLE 4.5. MEE Values for the Example

MEE1 = RSC1 - CVC1 = .333 - .375 = -.042
MEE2 = RSC2 - CVC2 = .750 - .625 = .125
MEE3 = RSC3 - CVC3 = .990 - .875 = .025

Those MEE values *should not* be simply added, because MEE1 and (RSC1) has not been computed on the same number of observations (here six) as MEE2 and (RSC2) has been (here four).

For this reason, the MEE values should be weighted by the RUCs, that is, by the *rates* of use of each confidence degree, presented in Table 4.6.

TABLE 4.6. Rates of Use of Confidence Degrees (RUCs) in the above example where NQ is 20 (see TABLE 4.2.).

RUC1 = 6/20 = .3
RUC2 = 4/20 = .2
RUC3 =10/20 = .5

The calibration for index PSYCA is computed as follows:

$$\psi CA = \sum_{i=1}^{nc} RUC_i \ (CVC_i - RSC_i) \quad \text{or} \quad PSYCA = \psi CA = \sum_{i=1}^{nc} RUC_i \ MEE_i$$

where nc = number of confidence degrees.

RUC_i = rate of use of confidence degree i.
RSC_i = rate of success with confidence degree i.
CVC_i = central value of confidence degree i.
MEE_i = mean error of estimation with confidence degree i.

This kind of formula was potentially contained in Brier (1950) and has been developed by Murphy (1972, 1973, and 1974).

A *negative* value of PSYCA is a symptom of underestimation whereas a *positive* value is an indication of overestimation.

Note that, since a half width of each zone is equal to .125, underestimation begins only when PSYCA < -.125 and overestimation begins only when PSYCA > .125.

THE MEASUREMENT OF REALISM (R)

In the above example, PSYCA is the result of a sum with one negative term and two positive terms, as it appears in Table 4.7.

TABLE 4.7. Details of Computation of PSYCA for the Given Example. Note that the left Term of the Sum is negative whereas the two other terms are positive.

$$\psi CA = [\,(.3)\ .\ (-.042)\,] + [\,(.2)\ .\ (.125)\,] + [\,(.5)\ .\ (.025)\,]$$

This example shows that a PSYCA could have a null value resulting from a compensation of negative values by positive ones.

The principle of the ψR index is to sum up the estimation errors whatever their sign (negative or positive).

Murphy (1973) suggests using the square of the MEEs in order to have a positive sum. This index of realism will be called PSYMR (M after Murphy):

$$PSYMR = \psi MR = \sum_{i=1}^{nc} RUC_i \ (MEE)^2$$

To measure what he calls the "appropriateness of confidence", Oskamp (1962) has suggested a formula where the *absolute values* of the MEE are summed. As suggested by Lefevre (1978), we called the resulting index PSYR (R after realism).

$$PSYR = \psi R = \sum_{i=1}^{nc} RUC_i \ |MEE_i|$$

Typical values computed for the above example are presented in Table 4.8.

TABLE 4.8. An Index of Calibration (PSYCA) and two Indexes of Realism (PSYMR and PSYR) for the Example presented in TABLE 4.2.

ψ CA = .024
ψ MR = .004
ψ R = .051

Since a half width of a confidence zone is .125, this student can be considered as realistic: his ψR is lower than .125.

Note that ψCA varies from -1 to +1 whereas ψR varies from 0 to 1.

Adams and Adams (1961) proposed a "mean absolute discrepancy score" (that will be referred to as ψAA):

$$\psi AA = PSYAA = \frac{\sum_{j=1}^{ne} \sqrt{NUC_j} - |\,CVC_j - RSC_j\,|}{\sum_{j=1}^{ne} \sqrt{NUC_j}}$$

THE MEASUREMENT OF EFFICIENCY IN THE USE OF CONFIDENCE DEGREES

The index of efficiency will be called PSYE. This mathematical index depends upon the tariffs whereas the previous (ψCO, ψCA, ψMR and ψR) did not.

It is based on the principle that the quality of self assessment can be measured only by referring to the correctness (objective value) of the response. For this reason, ψE could be called the *conditional ψ index*. Actually, it depends upon the number of correct responses (NCR) and the number of incorrect responess (NIR) to a test; these two values are known at the beginning of the computations.

Let us go back to the 20 question-test and to Table 4.2 where our student had answered 14 times correctly (NCR = 14) and six times incorrectly (NIR = 6). The first step for computing ψE is the drawing of the table of the possible distributions of confidence indexes. The vertical sums of such a table are already determined; in our example, they are 6 and 14 respectively (see Table 4.9). *Given such a situation*, the worst and the best, two extreme situations can be considered (Tables 4.10 and 4.11).

The constraints

	NIR	NCR
CO		
C1		
C2		
C3		
	6	14

TABLE 4.9. The Basic Conditional Data Matrix

The worst situation

	NIR	NCR
CO		14
C1		
C2		
C3	6	
	6	14

TABLE 4.10. The Worst Data Matrix

The best situation

	NIR	NCR
CO	6	
C1		
C2		
C3		14
	6	14

TABLE 4.11. The Best Data Matrix.

The score that will be computed for the worst distribution is the minimal possible score with confidence tarriffs. It will be referred to as the MINSCT. Here the value is 6 x TIC3 = 6 x -5 = -30. The test score that will be computed for the best distribution is the maximal possible score with confidence tariffs (MAXSCT). Here the value is 14 x TCC3 = 14 x 5 = 70. It can already be predicted, without any error, that the score (with a C t

scale) on the test (SCT) will be greater to or equal to -30 and lower or equal to 70. Acutally, SCT for our student (see Table 4.2) is equal to 52. The computation of this score is detailed in Table 4.12.

TABLE 4.12. Details of Computation of the SCT

$$(2 \times 3) + (3 \times 4) + (9 \times 5) = 6 + 12 + 45 = 63$$
$$(4 \times -1) + (1 \times -2) + (1 \times -5) = \underline{\quad -11}$$
$$52$$

The relative position of SCT on the range of the possible scores is considered as the efficiency index:

$$\text{PSYE} = \psi E = \frac{\text{SCT} - \text{MINSCT}}{\text{MAXSCT} - \text{MINSCT}}$$

Note that MINSCT value can be negative or positive. In our example, the ψEvalue is .82 (see details in Table 4.13).

TABLE 4.13. Details of Computation of ψE.

Note that $0 \leqslant \psi E \leqslant 1$.

$$\psi E = \frac{52 + 30}{70 + 30} = \frac{82}{100} = .82$$

A PROCEDURE FOR RAPID COMPUTATION OF ψ INDEXES

The above indexes are irksome for hand computation. The following procedure allows quick hand calculations, since five simplification principles have been adopted.

Five Simplification Principles

a. The instructions present eleven degrees of confidence, nine of them having a "round number" as a central value (CVC), expressed in integers (1, 2, 3, 4 ...) as shown in Figure 4.1.

b. The CVCs are multiplied by 10, so that only integers are used.

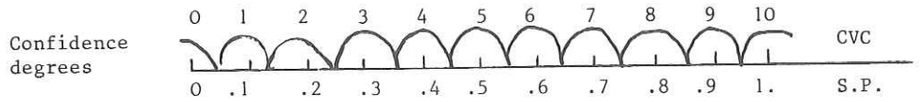


Fig. 4.1. Graphic representation of instructions offering ELEVEN degrees of confidence and rapid computation.

c. The new CVC0 is fixed at 0 whereas its real value is 0.025 and the new CVC10 is fixed at 10 whereas its real value is 0.975.
d. The number of questions is fixed at 10.
e. The tariffs are:
   TC = 10 (in the case of a correct response).
   TC = 0 (in the case of an incorrect response).

The new scale is referred to as the 10 t scale. No omission is allowed.

Computation Procedures

For a given item j and a given student, C is the chosen confidence index and S10Qj is his score for the item on the 10 t scale. The error of estimation for item j is:

$$EE_j = C_j - S10Q_j$$

When EE_j is negative, it is an indication of underestimation whereas it is an indication of overestimation when it is positive.

$$\psi CA = \frac{\sum_{j=1}^{10} C_j - SQ10Q_j}{100}$$

$$\psi R = \frac{\sum_{j=1}^{10} |C_j - S10Q_j|}{100}$$

Of course, ψCA is also equal to the difference between $\sum_{j=1}^{10} C_j$ and S10T (the score on the total), the difference being divided by 100:

$$\psi CA = \frac{(\sum_{j=1}^{10} C_j) - S10T}{100}$$

Table 4.14 displays an example of responses and use of confidence degrees with the instructions presented in Figure 4.1.

TABLE 4.14. An Example of Confidence Degrees, Question Scores and Error of Estimates to a 10 items Test.

Item number	1	2	3	4	5	6	7	8	9	10	Σ
Confidence index	5	4	5	6	9	7	7	9	6	7	65
S10 Q	10	0	0	10	10	0	10	10	0	0	50
C - S10 Q (or EE)	-5	4	5	-4	-1	7	-3	-1	6	7	15

Typical Results

In 1981, A. and M. Mathues developed for hospital nurses a course on the "artificial lung". Four groups (of about 20 students each) used various combinations of supports as programmed learning in a booklet, audio-visual aids, computer assisted instruction.

This last method (CAI) used the DOCEO system (Houziaux, 1965; Houziaux, 1972) including a special terminal (with audio-visual facilities) and an education-oriented language named LPC (in French: Language for Programming Conversational Processes) (Bartholome & Houziaux, 1979). This system has been used both in medical context (Lefevbre & Houziaux, 1969; Houziaux et al, 1978) and in school situations (Jamart et al, 1983; Leclercq, 1980).

In the Mathues experiment, each group received 40 questions before the training and the same 40 questions after. In addition to each answer, the students had to provide a confidence degree according to the instructions presented in Figure 4.1.

The FORTRAN computer program called ELEVEN (since there are 11 available confidence degrees) prints out, for each student:

a) the average "10 score" (S10T) to a text (minimum is 0 and maximum is 10).
b) The average confidence degree to a text (from 0 to 10).
c) The PSYCA value (calibration).
d) The PSYR value (realism).
e) The table presenting the rates of use (RU) of each confidence degree.
f) The table presenting the rates of success at each confidence degree.

The a, b, c, and d information for group 1 (24 students) is presented for the pretest in Table 4.15 and for the post-test in Table 4.16.

TABLE 4.15. Individual Pretest Results (S1OT) scores produced by program ELEVEN for Group 1 (24 students).

PRETEST GROUP 1

CODE	SCORE	CONF	PSYCA	PSYR
0101	2.25	1.27	-0.97	1.47
0102	1.50	1.67	0.17	1.32
0103	3.00	0.57	-2.42	2.88
0104	2.25	2.38	0.13	2.88
0105	4.50	4.92	0.42	3.47
0106	2.00	1.82	-0.17	1.67
0107	3.00	1.90	-1.10	2.75
0108	3.00	1.42	-1.57	2.97
0109	2.50	1.77	-0.72	2.13
0110	2.50	1.42	-1.07	1.82
0111	2.25	1.63	-0.63	1.13
0112	1.25	0.52	-0.72	1.17
0113	1.00	0.27	-0.72	1.02
0114	1.50	0.95	-0.55	1.25
0115	2.75	3.38	0.63	2.52
0116	1.50	1.55	0.05	1.50
0117	3.00	3.02	0.02	3.27
0118	2.00	1.20	-0.80	1.35
0119	1.75	1.47	-0.27	1.17
0120	2.00	0.35	-1.65	2.10
0121	1.00	0.47	-0.52	0.52
0122	0.75	0.27	-0.47	0.72
0123	3.00	2.02	-0.97	2.52
0124	1.25	0.75	-0.50	0.65
MOY=	2.15	1.54	-0.60	1.85

TABLE 4.16. Individual Post-test Results (S1OT scores) produced by program ELEVEN for Group 1 (24 students).

POST-TEST GROUP 1

CODE	SCORE	CONF	PSYCA	PSYR
0101	8.75	9.25	0.50	1.80
0102	9.00	9.92	0.92	0.92
0103	9.50	8.00	-1.50	2.00
0104	9.25	10.00	0.75	0.75
0105	9.75	10.00	0.25	0.25
0106	9.00	9.95	0.95	1.05
0107	9.25	10.00	0.75	0.75
0108	9.25	8.17	-1.07	2.87
0109	9.75	9.30	-0.45	0.75
0110	9.75	9.57	-0.17	0.63
0111	9.75	9.82	0.07	0.42
0112	9.25	9.70	0.45	1.00
0113	8.75	9.30	0.55	1.30
0114	9.75	9.80	0.05	0.20
0115	9.00	9.35	0.35	0.90
0116	9.00	8.97	-0.02	1.38
0117	9.25	9.30	0.05	1.25
0118	10.00	9.72	-0.27	0.27
0119	10.00	10.00	0.0	0.0
0120	9.25	7.92	-1.32	2.42
0121	9.00	6.07	-2.92	3.82
0122	10.00	9.25	-0.75	0.75
0123	10.00	10.00	0.0	0.0
0124	9.50	8.57	-0.92	1.22
MOY=	9.41	9.25	-0.16	1.08

As expected, the average score on the pretest is low (2.15) whereas it is high at post-test (9.41). The average confidence degrees are close to these values: repectively 1.54 (instead of 2.15) and 9.25 (instead of 9.41), producing low PSYCA average values (-0.60 and - 0.16), meaning good calibration.

It can be noted that PSYR (realism) average value drops from 1.85 on the pretest to 1.08 on the post-test. This has been observed for the four groups (see Table 4.17).

Figure 4.2 presents graphically the information given in a, b, and c columns of Tables 4.15 and 4.16. The points appearing at the right hand side (top) present post-test data.

The student indicated by an arrow (student 121) appears to underestimate seriously (his PSYCA is equal to -2.92).

TABLE 4.17. Average Pretest and Post-test Values of PSYR for the four Experimental Groups in MATHUES' Experiment.

	Group 1	Group 2	Group 3	Group 4
Pretest	1.85	1.51	0.77	1.04
Post-test	1.08	0.64	0.58	0.65

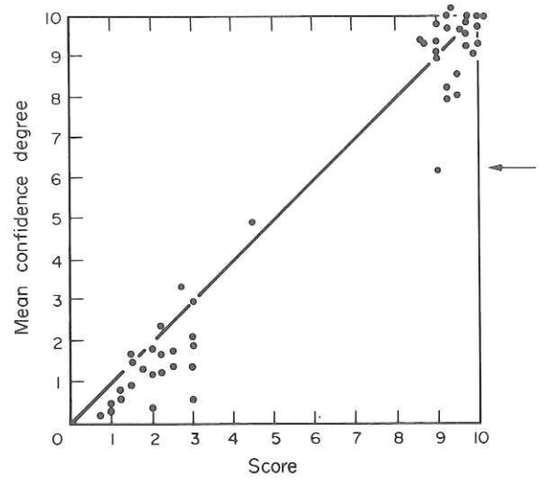


Fig. 4.2 Plot of Individual Mean Confidence Degrees against Individual Score.

Tables 4.18 and 4.19 present detailed data (information e and f of program ELEVEN output) that can explain how such an underestimation happened.

From Table 4.18, it appears that student 121 used a great amount of low indices (0, 1, and 2), i.e., in 26% of the times, whereas the responses were correct (rate of success = 100%), as appears in Table 4.19.

The distance between each point and the diagonal appearing in the plotting of Figure 4.2 is not indicative of the PSYR (realism values) but of the PSYCA. For instance, student 117 at pretest (Table 4.15) has a very good PSYCA index (0.02) whereas he has a bad PSYR index (3.27).

TABLE 4.18. Rates of Use for 24 Students (of Group 1), of each of the 11 Confidence Degrees at Post-test.

	0	1	2	3	4	5	6	7	8	9	10
0101	0	0	0	0	0	0	0	4	29	0	64
0102	0	0	0	0	0	0	0	0	2	2	94
0103	0	0	0	0	2	17	4	9	19	7	37
0104	0	0	0	0	0	0	0	0	0	0	100
0105	0	0	0	0	0	0	0	0	0	0	100
0106	0	0	0	0	0	0	0	0	2	0	97
0107	0	0	0	0	0	0	0	0	0	0	100
0108	2	0	0	0	0	9	2	19	17	2	44
0109	0	0	0	0	0	2	7	2	4	9	72
0110	0	0	0	0	0	0	4	0	0	22	72
0111	0	0	0	0	0	0	2	2	0	0	94
0112	0	0	0	0	0	0	0	2	4	12	79
0113	2	0	0	0	0	0	0	9	4	4	77
0114	0	0	0	0	0	2	0	0	2	2	92
0115	2	0	0	0	0	4	0	4	0	0	87
0116	0	0	0	0	2	9	0	2	14	0	69
0117	0	0	0	0	0	0	0	17	7	2	72
0118	0	0	0	0	0	0	0	2	4	9	82
0119	0	0	0	0	0	0	0	0	0	0	100
0120	0	0	0	4	0	0	19	7	25	19	22
0121	9	9	4	4	0	22	0	4	0	0	42
0122	0	0	0	0	0	7	0	12	0	0	79
0123	0	0	0	0	0	0	0	0	0	0	100
0124	0	2	0	0	0	12	2	4	9	12	54

TABLE 4.19. Rates of Success for 24 Students (of Group 1) of each of the 11 Confidence degrees at Post-test.

	0	1	2	3	4	5	6	7	8	9	10
0101	0	0	0	0	0	0	0	100	83	0	88
0102	0	0	0	0	0	0	0	0	0	0	94
0103	0	0	0	0	0	100	50	100	100	100	100
0104	0	0	0	0	0	0	0	0	0	0	92
0105	0	0	0	0	0	0	0	0	0	0	97
0106	0	0	0	0	0	0	0	0	100	0	89
0107	0	0	0	0	0	0	0	0	0	0	92
0108	100	0	0	0	0	75	100	87	85	100	100
0109	0	0	0	0	0	100	66	100	100	100	100
0110	0	0	0	0	0	0	100	0	0	88	100
0111	0	0	0	0	0	0	100	100	0	0	97
0112	0	0	0	0	0	0	0	100	100	80	93
0113	0	0	0	0	0	0	0	0	75	100	100
0114	0	0	0	0	0	0	0	0	100	100	100
0115	0	0	0	0	0	50	0	100	0	0	94
0116	0	0	0	0	0	75	0	100	83	0	96
0117	0	0	0	0	0	0	0	85	100	0	96
0118	0	0	0	0	0	0	0	100	100	100	100
0119	0	0	0	0	0	0	0	0	0	0	100
0120	0	0	0	100	0	0	87	100	80	100	100
0121	100	100	100	50	0	66	0	100	0	0	100
0122	0	0	0	0	0	100	0	100	0	0	100
0123	0	0	0	0	0	0	0	0	0	0	100
0124	0	0	0	0	0	80	100	100	100	100	100

Uses of such indices in school settings would imply the presentation of those results through appropriate *phrasing*. For this purpose, computer programs could help a lot.

THE INTERPRETATION OF VARIOUS SCALES OF SCORES

Let us recall that the simple score to a test (SST, that is, the score computed with the S t scale), or the score corrected for guessing (SGT) are *measures* of ability and that various ψ indexes are *measures* of realism, calibration, or acuity of self assessment. Contrary to this SCT, a score on a test, computed from a C t scale or a C t matrix is *not a measure* of ability, but a payoff, a reinforcement.

Why Use SCT Scores?

The usefulness of SCT scores can be questioned since the teacher already possesses relevant information on the two interesting variables: the student's ability (SST or SGT scores) and the student's self assessment (ψ indexes). Despite this, there are at least two reasons for using SCT scores in educational settings. First, since the matrices of tariffs are computed to constrain the student in telling the truth, those tariffs must actually be explained to the students who must experience the consequences of their choices and, consequently, adapt their behavior. Second, by using SCT scores that are a mixture of measures of ability and self-assessment, the teacher stresses the point that social usefulness (or relevance) of ability is not limited to the raw quantity of knowledge but includes the ability to handle it properly.

The Nature of SCT Scores

If SCT scores are used, new problems arise, because SCT refer to an uncommon scale and the potential users (teachers, parents, students) usually are familiar with quite a different type of scale.

For instance, if the maximum score is 20, a score lower than 10 is considered a bad result and frequently is associated with "failure". In this familiar mental scale, no negative scores exist and scores do not exceed 20 (the maximum).

With correction for guessing procedures (SGT) scores, negative values are possible but people usually have difficulties in interpreting such scores. With the SCT scores, the same problem arises. Of course, it could be explained to parents that test scores are metric values of an interval scale but not of a ratio scale where 0 means "absence of the property" (as is the case with weight, height, duration, etc.): test scores have no absolute zero point.

Interpreting Negative Scores

How can we interpret a negative score? Let us recall that in any scale where TO (tariff of omission) is 0, a student that omits all items will receive a zero score. This student (and his score) can be used as a reference for "someone that knows nothing on the tested topic". So, a negative score could be interpreted as "worse knowledge than no knowledge at all". This is a quasi moral interpretation, but in crucial domains (medicine, rescue, chemistry, ...) it could be phrased "a more dangerous knowledge than no knowledge at all".

Increasing Requirements

Under certain conditions negative scores might be understood and accepted but many people would still be shocked by unusual frequency of low scores (lower than 10/20). This is caused not only by the presence of negative values among the tariffs but also by the introduction of a new requirement about performance. Perfection is no more providing all the correct answers, but providing them with the highest degree of confidence. This, in turn, makes SCT scores incomparable with traditional scores; consequently, the SCTs must be adapted. The following section explains how such a transformation could be made.

Such an adaptation is necessary since students will not accept a new scale or a new procedure that appears to handicap them compared with another scale or procedure; and such a reaction is perfectly legitimate.

The Problem of Severity

Let us refer to MMAX to designate the mathematical maximum of a test score with a given tariff. We already mentioned that MMAX was more difficult to reach using SCT than when using SST (simple scale of tariff).

So it is reasonable to fix an arbitrary RMAX (that is a reference maximum) lower than MMAX in such a way that someone who obtains a score equal or superior to RMAX gets the top award or top marks in the school (i.e., the final score that is communicated to parents). We shall refer to this final maximum as the FMAX. (for instance 10/10 or 20/20 or 100/100).

The value of RMAX depends upon the teacher's severity. Let us illustrate this using tariffs that appear in Table 4.20 (equivalent to Table 4.1).

TABLE 4.20 . Typical L Matrix.

Confidence degree	TI	TC
0	0	0
1	-1	+3
2	-2	+4
3	-5	+5

MMAX is obtained by NQ x TCmax (here NQ x +5) RMAX is obtained by NQ x SEV (SEV = severity). Experience indicates that, in secondary schools, SEV should vary between +3 and +4. The maximum severity is SEV = TCmax (then MMAX = RMAX).

If SEV is fixed to +4, RMAX would be obtained by a student that gives all correct responses with a confidence degree of 2.

For 20 items, RMAX would be equal to 80 with SEV being equal to +4. Of course, some students could obtain a SCT score equal to (or even higher than) 80, even with some omissions or errors, and receive a final score equal to FMAX (e.g., 100%). This property of the severity correction should lower the student's anxiety when answering a test: an omission or even an error no longer means losing the chance of being able to achieve a maximal final score or school mark.

The final test score (with confidence tariffs) is computed by the following formula:

$$FSCT = \frac{SCT \times FMAX}{RMAX} = \frac{SCT \times FMAX}{NQ \times SEV}$$

An example of this transformation principle is given in Table 4.21. The experiment took place in 1971, at the Belgian Air Force Technical School where nine technicians were tested after a two month intensive course on the MIRAGE IV Army airplane.

The test contained 25 questions, the tariff matrix was equivalent to the one presented in Table 4.16, and the severity was equivalent to +4, so RMAX was equal to 100 and MMAX to 125. The final score was expressed on a scale with an FMAX equal to 20.

Here, $$FSCT = \frac{SCT \times 20}{25 \times 4} = \frac{SCT \times 20}{100} = \frac{SCT}{5}$$

TABLE 4.21. Tests Scores (SCT) and Final Test Scores (FSCT)

Confidence d. / TARIFFS	Incorrect response				Correct response					Written on report (school marks)
	3	2	1	0	1	2	3			
NAMES	-5	-2	-1	0	+3	+4	+5	SCT	FSCT	
1. DO ...	0	0	0	2	0	0	23	115	23,0	20
2. LE ...	0	0	0	0	4	11	9	101	20,2	20
3. NE ...	0	0	1	0	0	6	19	118	23,6	20
4. CA ...	0	1	0	1	0	14	9	99	19,8	19,8
5. HA ...	0	0	0	1	0	10	14	110	22,0	20
6. LO ...	1	0	0	2	1	6	15	97	19,4	19,4
7. MO ...	0	0	0	0	0	10	15	115	23,0	20
8. VA ...	0	0	0	1	0	17	7	103	20,6	20
9. WI ...	0	1	0	0	0	4	20	114	22,8	20
TOTAL	1	2	1	7	5	73	73	131		

It can be observed that seven technicians out of nine reached (and exceeded) RMAX. So, FMAX (here 20) was obtained by nearly everyone. This is a normal result for highly qualified adults who have been exposed to 82 instructional hours on a crucial topic.

The above mathematical transformations may appear artificial, but it would be unrealistic to ignore the comparability problem.

There is still a great deal to be discovered about human characteristics and self evaluation. For example, into which limits should the estimations be considered as acceptable? Further research is necessary to examine this kind of problem since we are just beginning to gather reliable data in this field. The problems of interpretation are a central concern; they should be refined in the future, but we must stick to the assumption that "it is only subjective probability that can give an objective meaning to every response and scoring method" (De Finetti, 1965, p. 111).

CHAPTER 5

THE STABILITY AND THE ACUITY OF CONFIDENCE DEGREES

Whereas the stability and acuity are individual characteristics, their average values on populations are of great interest to help researchers and teachers in devising or selecting the most suitable test instructions. The experiments described below were undertaken for this purpose.

THE CLASSICAL APPROACHES

Stability

Instead of the classical word "reliability", we shall use the term "stability" to express the degree of unvariability of confidence degrees given by an individual across repeated questioning. In cognitive testing, it is almost impossible to administer the same test twice to the same subject, under the same conditions since learning will have occurred between test administrations. Consequently, in classical test theory, reliability is often estimated by an artifact, i.e., internal consistency (split-half method or K.R. Formula). Various kinds of "split-half" methods could be developed to cope with the problem of reliability in the classical way (the internal consistency concept, the part-whole correlations, and the Spearman-Brown type formulas), adapted to subjective probabilities. However, a more direct approach will be presented, that is a test-retest method. The strengths and weaknesses of this approach will be discussed after the analysis of the results.

Acuity

Acuity is an individual's characteristic and is sometimes called *sensitivity*, *sharpness*, or *resolution*. Lichtenstein and others (1977, p. 279) define it as "the ability of the assessor to sort the event into subcategories for which the hit rate is maximally different from the overall hit rate."

Murphy's formula (1973) is the mathematical version of this definition:

$$\psi MA = \sum_{j=1}^{nc} RUC_j \, (RSC_j - RST)^2$$

where ψ_{MA} = MURPHY's index of acuity (or resolution)

RUC_j = rate of use of confidence degree j

RSC_j = rate of success of confidence degree j

RST = rate of success for the total test

nc = number of confidence degrees.

A derived formula could be:

$$\psi A = \sum_{j=1}^{nc} RUC_j \left| RSC_j - RST \right|$$

Murphy (1972) suggested an individual "overall calibration index" using the formula which he called a "special scalar partition":

$$\psi M = RST \, (1 - RST) + \psi_{MR} - \psi_{MA}$$

AN EXPERIMENTAL APPROACH

In the following experiment, two problems (stability and acuity) are studied in the same experimental design based on a "confidence guessing game" (CGGame).

The Confidence Guessing Game (CGGame)

The confidence guessing game presented here is directly inspired by *Shannon's guessing game* (1951) in which the subject has to predict successively each letter of an English text. There are only 27 possible answers (each of the 26 letters, plus the "blank" for the spaces, points, etc.). In Shannon's method, when an answer is wrong, the subject has to give other letters until he finds the correct one. In this way, the experimenter can indicate below each letter the number of trials needed until the correct answer was found.

Attneave (1959, p. 30) has shown (see Table 5.1) a typical result obtained by Shannon. It appears that the last letters of a long word (e.g., Dramatically) are quickly discovered (on the first trial).

TABLE 5.1. Numbers of Trials a Student needed to discover the Correct Letter in Shannon's Game.

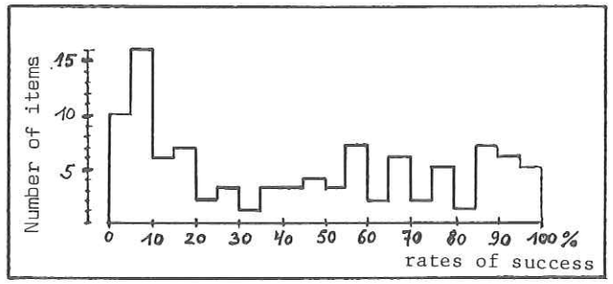
T	H	E	R	E		I	S		N	O		R	E	V	E	R	S	E		O	N		A		M	O	T	O	R	C	Y	C	L	E
1	1	1	5	1	1	2	1	1	2	1	1	15	1	17	1	1	1	4	1	3	2	1		2	2	7	1	1	1	1	4	1	1	1	1	1
A		F	R	I	E	N	D		O	F		M	I	N	E		F	O	U	N	D		T	H	I	S		O	U	T						
3	1	8	6	1	3	1	1	1	1	1	1	1	1	1	1	4	6	1	1	1	1	1	1	1	2	1	1	1	1	1	1						
R	A	T	H	E	R		D	R	A	M	A	T	I	C	A	L	L	Y			T	H	E		O	T	H	E	R		D	A	Y			
4	1	1	1	1	1	1	15	1	1	1	1	1	1	1	1	1	1	1			6	1	1	1	1	1	1	1	1	1	1	1	1	1			

The principles and rules of the CGGame are as follows:

The items. A long text (about 3 pages) is chosen from a book and punched onto cards. The odd cards are printed whereas the even ones are not. The subjects are asked to predict the first letter of the omitted sequence, without the use of a dictionary (since this could possibly produce a probabilistic strategy based on letter frequencies in the dictionary).

If the experimenter follows this procedure blindly, he will obtain too many easy questions (the beginnings of many words indicate unambiguously the following letters). So, cut-off points must be modified slightly, in order to obtain items of various difficulties (ideally, a rectangular distribution with an average mean of .50). Table 5.2 shows the distribution of the facility indexes for the 100 questions in the experiment.

TABLE 5.2. Distribution of the percent of success for each of the 100 items of the experiment. (The ideal would have been a rectangular distribution.)



The responses. Table 5.3 presents an example of such an item. The subjects were requested to write the next letter (here the letter L) and to "circle" a subjective probability on each of the three scales.

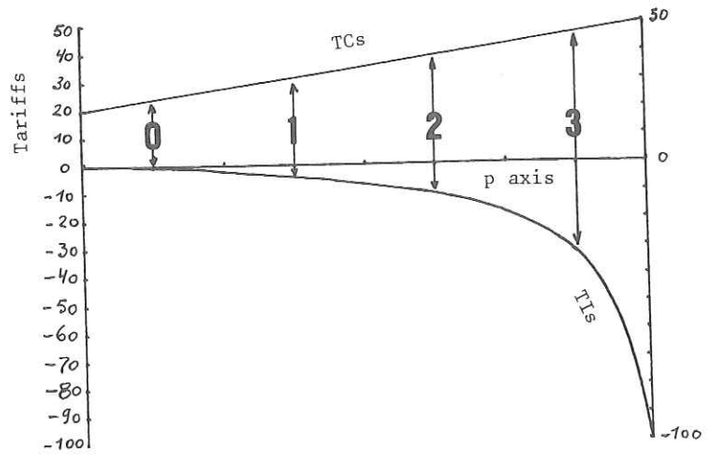
TABLE 5.3. A typical example of the Confidence Guessing Game. (Here, the correct answer is L, since the truncated text is "The magical number seven plus or minus two.")

THE MAGICAL NUMBER SEVEN P

scale 1 → -----A----/----B----/----C----/----D----
scale 2 → 0 1 2 3 4 5 6 7 8 9
scale 3 → |...|...|...|...|...|...|...|...|...|...|
0 10 20 30 40 50 60 70 80 90 100

The tariffs. The subjects were told that "points would be given in such a way that in order to maximize their total score they should not bias their subjective estimate" (i.e., they should tell the truth). The table of wins and losses (tariffs) for given probabilities, as well as their plotting, were presented to the students (see Table 5.4). The maximal score is +50 (TC for confidence degree 100 on scale 3 of Table 5.3) whereas the minimal score is -100 (TI for confidence degree 100 on scale 3 of Table 5.3).

TABLE 5.4. TCs and TIs Tariff Curves for Scales 1, 2 and 3 of Table 5.3.



Advantages of the CGGame

This confidence guessing game (CGGame) contains various advantages:

a) The CGGame is close to the school situation:

- it is possible to decide whether an individuals answer is correct or not;
- it is impossible to "compute" objectively the subject's "internal state of uncertainty".

b) The CGGame is suitable for experiments:

- possible content is infinite, in any language, easily available, for any age or category of interest;
- the difficulty level of the test can be easily adapted (the experimenter can choose which letters will be suppressed);
- instructions are the same for all the items and easy to understand;
- listing by computer is quick and inexpensive;
- correction for student feedback and experimental research is easily done on the computer (an answer = one position on a punched card);
- the task is self-motivating for the subject;
- it can be modified in order to deal with Bayesian probability theory (see Chapter 6).

c) The CGGame is suitable in a test-retest setting:

- the subjects have to provide so many confidence indexes (100) that they are unlikely to remember them.
- unless they discover from which book the text has been extracted (none of our subjects succeeded in discovering it), no information on the answers can be gained between the test and the retest.

The Experimental Setting

The experiment was conducted in three steps:

a) The "guessing game" and the scoring rules were explained to about 300 high school teachers. A dry run was conducted with 5 items; each participant recieved the correct answers and his score a few days after (by mail). The results appeared on computer listings, and comments were given such as each participant's rank order, or overall tendency to overestimate or to underestimate.

b) In the experiment itself (test), subjects were requested to answer 100 items and to assign to each answer a subjective probability (SP) of correctness. SP had to be expressed on three different scales:

- Scale A: four possible confidence degrees (25% each);
- Scale B: ten possible confidence degrees (10% each);
- Scale C: forty possible confidence degrees (2.5% each).

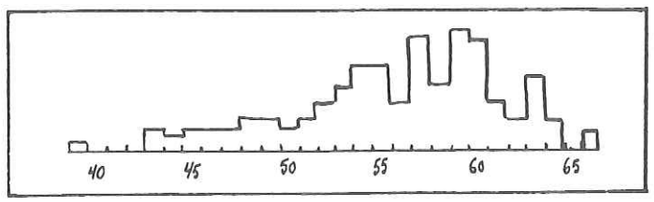
c) One month later (retest), subjects received the same questions and their *answer* (they were not allowed to change the *answers*) but did not receive their previous SPs. Subjects were requested to give their SPs again.

d) Furthermore, on retest, subjects were invited to describe the way in which they chose a given degree of confidence.

The General Results

The whole test-retest procedure was completed by 124 subjects. Three subjects having awkwardly low scores were discarded. As can be seen in Table 5.5, the distribution of the 121 simple scores (SST = number of correct letters) is close to a normal distribution, with extremes being 39 and 66, with a mean of 56.

TABLE 5.5.



For 78 subjects, confidence scores (SCT: score computed with the confidence tariffs) were better on the first test than on the retest; for 42 subjects, the contrary occurred. Only one subject had the same score for both testings. The loss in efficiency on the retest could possibly be attributed to boredom having to answer 100 questions yet again.

RESULTS CONCERNING STABILITY

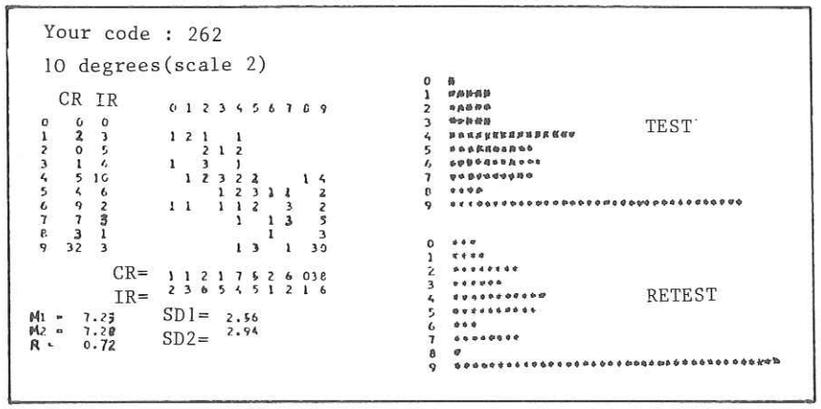
The Basic Data

For each subject, a FORTRAN computer program prints:

- The scatter plot of the 100 confidence degrees used on the test and on the retest;
- The two means (M1 and M2) and standard deviations (SD1 and SD2);
- The Bravais Pearson correlation coefficient (R);
- The histogram of the 100 degrees of confidence on the test;
- the histogram of the 100 degrees of confidence on the retest;
- An example of the printout is given in Table 5.6.

Moreover, the Kolmogorov-Smirnov (D max) coefficient for differences between the two distributions has been computed (here the null hypothesis had to be rejected at the .05 level for a D max greater than 0.192).

TABLE 5.6. Printout for an Individual Participant
where CR = Number of correct responses
and IR = Number of incorrect responses.



Analysis of the Distributions

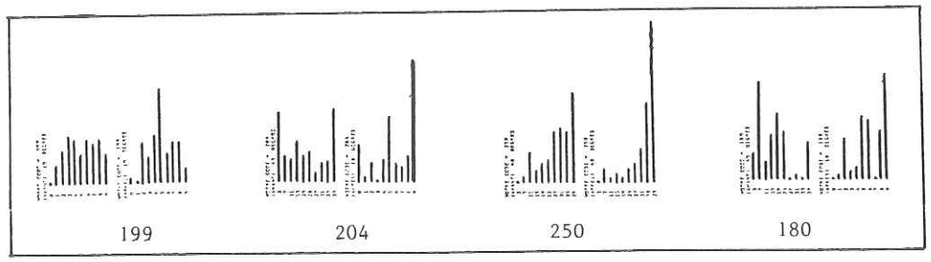
In Table 5.6, subject 262 has two equivalent distributions (D max < 0.192). Table 5.7 presents examples of four different reasons causing the rejection of the null hypothesis:

1: modes differ on test and retest (subject 199)
2: skewnesses differ (subject 204)
3: kurtosises differ (subject 250)
4: modes and skewnesses differ (subject 180) whereas kurtosis does not differ.

From 121 subjects, the Kolmogorov-Smirnov hypothesis was not rejected in 55 cases. This may be a result of the following weaknesses in the method of testing:

1. Subjects can guess blindly in order to finish quickly. Winning points is not a "satisfying state of affairs" for everyone; money might have been more "operant" and stimulating. This weakness is not the most important since only three "fantastic results" have been

TABLE 5.7. Four examples for whom the distribution of confidence degrees on the re-test does not fit the distribution on the pre-test.



observed. An ideal setting might be conducting the questioning and answering through a computer terminal that displays immediately the amount of dollars won after the whole test (or after a series of 20 questions).

2. In the retest, subjects can change their hypotheses about the correct letter (they find new possible words and forget possible words which they had considered before, etc.). New hypotheses may make the given (and unchangeable) answer more or less plausible.

3. At the retest, subjects can also change their strategy. It is well known that, when placed for the first time in a confidence marking situation, subjects use various strategies which have been described in Chapter 3. This effect has been reduced by the dry run on 5 items.

Analysis of the Correlations

Since 121 students used Scale A (the rawest, with 4 degrees) on the test and on the retest, 121 correlation coefficients can be computed. Their median value is 0.56, and the extremes 0.20 and 0.80. No significant differences appear between median correlation values for the rawest scales (scale 4 - scale 4), the 10/10 (intermediate scales) and the 40/40 (most accurate scales).

RESULTS CONCERNING ACUITY

The Basic Data

A "replication histogram" has been built for each degree of confidence of each scale (4 histograms for the scale A, 10 histograms for the scale B, 40 histograms for the scale C).

The replication histogram of a given degree (say X) is established according to the following principles (see Figure 5.1).

- The various degrees are placed on the horizontal line.
- The height of each rectangle expresses the number of times (here the percentages) that each degree (Z) has been used at the very place where degree X was used in the other test (test or retest). In fact, the percentages are corrected by the relative rate of use of the given Z degree. Theoretically, when degree X has been used in a test, the same degree should be used in the other (that is Z should be equal to X). Actually, when degree X has been used in a test, degrees close to degree X are used in the other test (and degree X is the most used of all of them).

In order to make the graphs clear, the tops of the histogram rectangles have been joined. Only the curves are presented. Figures 5.1 and 5.2 present the replication curves computed from 40 subjects for the 4 degrees of the rawest scale (A).

Analysis of the Replication Histograms

As can be seen, the top of the curve for degree X is X. Our adult untrained subjects seem to have had no problem in dealing with 4 degrees (scale A) and we suspect that they can handle more sophisticated scales.

Figures 5.3 to 5.7 present the 10 replication curves for the B scale from the results of 20 selected individuals. It appears that the mode for X is X except for confidence degrees 7 (mode is 6), 3 (mode is 2), confidence degree 6 (mode is 7) and confidence degree 4 (mode is 3).

This "overlap of some degrees" seems to indicate that 10 degrees is too much, either for untrained individuals or for this kind of work. Again, the calibration curve (see Figure 5.8) shows the ambiguity between degrees 3 and 4 and between degrees 5 and 6.

It would appear as if people had best acuity (accuracy) at the extremes of the scale, since degree 0, 1, 8 and 9 are never confusing. These results indicated that less than 10 degrees should have been used, since our

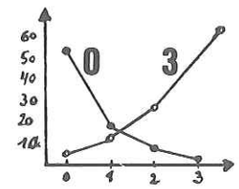


Fig. 5.1. Replication histograms of confidence degrees 0 and 3.

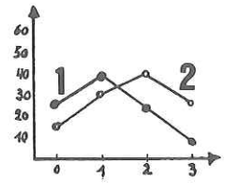


Fig. 5.2. Replication histograms of confidence degrees 1 and 2.

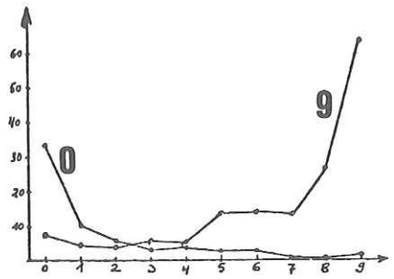


Fig. 5.3.

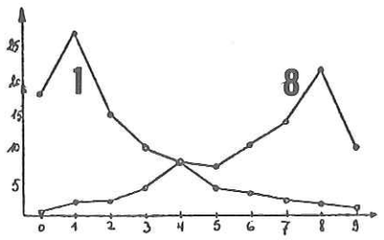


Fig. 5.4.

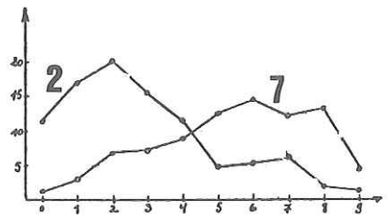


Fig. 5.5

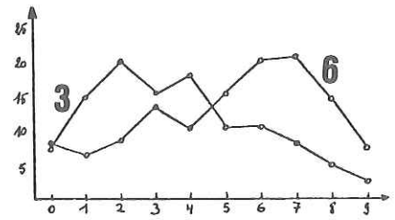


Fig. 5.6.

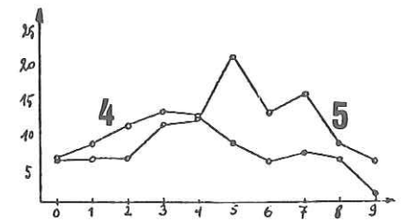


Fig. 5.7.

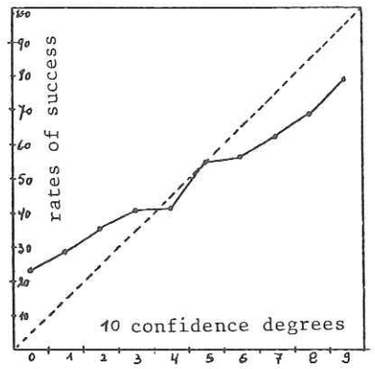


Fig. 5.8. Calibration curve constituted by the rates of successes for each value of scale 2.

subjects had difficulties in handling 10 degrees. With 40 degrees, it becomes impossible to build replication curves, because some degrees have too few data as can be seen from Figure 5.9 (the numbers of uses of each of the 40 degrees). Figure 5.10 presents the general calibration-curve, where reliable values (computed on sufficient data) are represented by dark dots.

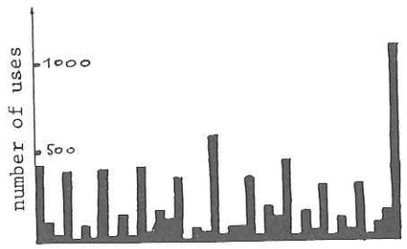


Fig. 5.9. The number of uses of each of the 40 confidence degrees of scale 3.

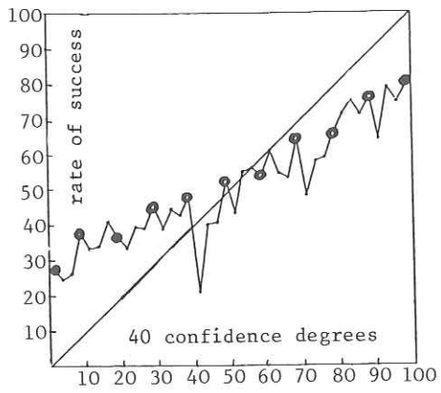


Fig. 5.10. General calibration curve generated by the rates of success for each value of scale 3.

Figure 5.9 indicates clearly that when given the choice among 40 degrees, people spontaneously choose only 11 of them. This does not necessarily mean that they can discriminate among the eleven chosen degrees. This result could be partly caused by the spatial disposition of the scale on the paper (see Table 5.3). Space for response could have been judged as too narrow. Further experiments should check this point.

DATA FROM SUBJECTS INTROSPECTION

In order to know how they choose a given degree of confidence, the subjects were presented a questionnaire (see three of the items hereafter). Moreover, the subjects were invited to write down any additional comments they wanted.

Question 1 : "When answering on scale 1 (with 4 degrees), 2 (with 10 degrees) and 3 (with 40 degrees), in which sequence do you use them?"

4 then 10 then 40
or 40 then 10 then 4
or 10 then 4 then 40
etc.

The 135 subjects who answered stated that they behaved as reported in Table 5.8.

TABLE 5.8. Number of Participants stating they have followed a Given Sequence in using the three Scales.

Starting with scale 1 :	4 then 10 then 40	: 56	
	4 then (10 or 40)	: 9	68
	4 then 40 then 10	: 3	
Starting with scale 2 :	10 then 40 then 4	: 15	
	10 then 4 then 40	: 11	32
	10 then (40 or 4)	: 6	
Starting with scale 3 :	40 then 10 then 4	: 18	
	40 then 4 then 10	: 10	35
	40 then (10 or 4)	: 7	
	Total		135

The great popularity of the 4 then 10 then 40 sequence could be caused by the typographical presentation of the scales (see Table 5.3): people are used to reading from top to bottom, not vice-versa. This point should be checked in other experiments.

Question 1 also allowed open responses to the check "I proceed another way." Here is a sample of four interesting answers to this question.

- When I am perfectly confident, I start from 100%; when I am not confident at all, I start from 0%.
- I select a given place on the line (from 0 to 100) without worrying about the numeric scale.
- For me, there are 3 situations: sure, doubtful (50%), not sure at all (0 to 10%). My "sure" responses are subsequently divided into

"perfectly sure (i.e., 100%)", "very likely (i.e., 95%)", and "likely (i.e., 80%)".
- I first gave the answers for the items I was perfectly sure of (i.e., 100%), and then I tried the rest of the items.

The Second Question

Suggested alternatives and percentages of choices are presented in Table 5.9.

TABLE 5.9. Percentages of Choices of the five Alternative suggested in question 2 of the Questionnaire.

Question 2 : "The choice of my degree of confidence depends on ...

% of response		nature of response
73 %	28 % ...	my SP (Subjective Probability) and the risk, but more on my SP (Subjective probability).
	26 % ...	my SP only.
	19 % ...	my SP and the risk, at the same level.
	18 % ...	the risk only.
	9 % ...	my SP and the risk, but more on the risk.

The popularity of the three first propositions is a good indicator for the validity of the whole procedure. Some subjects noted that their strategy depends upon the situation: if they had to risk their lives, things might change.

The Third Question

The answers are presented in decreasing order of frequency in Table 5.10.

TABLE 5.10. The most popular responses to question 3.

Question 3 : "My ideal number of degrees is ... (open ended question)".

Ideal scale provided by participants	Number of subjects (i.e. of participants)
10 degrees	= 41
20 degrees	= 13
40 degrees	= 8
100 degrees	= 4
10 to 20 degrees	= 3
4 degrees	= 3
50 degrees	= 2
5 degrees	= 1
6 degrees	= 1
7 degrees	= 1
8 degrees	= 1
9 to 10 degrees	= 1
4 to 10 degrees	= 1

A subject explained that he had chosen the 20 degrees scale because he had been used to it in school.

Some subjects who chose 10 degrees (the B Scale) as the ideal scale gave as comments that additional degrees should be necessary.

Here are two suggestions:

- "With the possibility of rating + and - for each degree" (for example 7 + and 7 -); this comes close to the 20 degrees scale.

- "With the possibility of using some special intermediate values" (as 25%, 75%, 95%).

Other interesting comments were made, such as the following:

- "My ideal scale is 10 degrees because there were 100 questions; with only 5 questions, 40 degrees would have been perfect."

- "Why not a continuous confidence marking (for example, the student should be allowed to answer 39%)?"

- "This game can be learned and subjects could improve their ability."

- "I would prefer to give several answers and give a probability to each of them." (This teacher was a linguist.)

In general, subjects pointed out the importance of the number and the kind of items, of the situation (real consequences or not), of the person (if he is "word-minded" or not) and the content of the text.

CONCLUSIONS OF THE APPROACH

To construct an ideal test-retest situation is difficult. To ensure the same level of uncertainty on the retest, subjects must not reformulate the hypotheses about the correct answer, change the hypotheses and, consequently, the probabilities. We have not succeeded in building a situation where subjective probabilities are expressed given fixed hypotheses. Such a conditional game should be developed.

From our Confidence Guessing Game (CGGame), it has been possible to observe reasonable test-retest stability. As for acuity, objective data as well as subjects' opinions showed that for this kind of game with adults, the optimal number of degrees was 10 or just below. This result corroborates

Miller's opinion that our spontaneous acuity in perceptual domains is "the magical number seven, plus or minus two." Moreover, we can formulate the hypothesis that our sensitivity (or acuity) is better at some portions of the probability axes (the extremes) thus supporting Edwards' procedure (1967) of using odds having logarithmic properties. (See chapter 2).

All of these observations are of interest for constructing an optimal scale to be used in school settings. It is also clear that, for practical reasons, we should only use a few degrees; but how many exactly and where on the axis of probabilities? It is hoped that the study reported above will inspire new experiments to help answer such questions.

CHAPTER 6

CONFIDENCE TESTING AND EDUCATIONAL RESEARCH

The use of partial knowledge coupled with good theories and techniques could be of great help to educational research. Research studies using partial knowledge have been undertaken in psychiatry (Been, 1970), in production planning (Kidd, 1970), in social psychology (McNeel & Messick, 1970; Lovie & Davies, 1970), and in meteorology (Murhpy, 1967-70; Epstein, 1967-69; Winkler, 1967-70).

Let us consider what might be done in education. Fundamental problems such as the informative power of a given event for a given individual, the study of humans as information processors and models of mental measurements will be considered. Four questions will serve as the starting points for our discussion:

- need for accuracy in educational research;
- relations between confidence testing and the Rasch Model;
- relations between confidence degrees, verifying behaviors, and performance;
- revision of confidence degrees caused by reception of information.

These problems will be discussed separately, but they are, of course, closely related. We are strongly convinced that the use of confidence testing and specially subjective probabilities will shed light on various important current problems in educational research.

THE NEED FOR ACCURACY IN EDUCATIONAL RESEARCH

In the following example, the classical researcher will be contrasted with the SP researcher who uses "subjective probabilities". The former, applying rigidly the concepts of behaviorism, will ask one response from the student, without taking into account expressions of doubt, confidence, etc. The latter will collect more detailed (and more subjective) data.

Let us consider the following (open ended) question presented to a Belgian student.

What was the political status of Malaysia in 1939?

Let us suppose that the individual who has to answer this item does not know the correct answer, and that he *considers possible answers such as*: Dutch

colony, English colony, Independent state, French colony, and Japanese colony. He can even do more than list or order these possible solutions: he can attribute to each a subjective probabilty of being the correct answer, as the SP researcher requested. Let us suppose that the answers are as in Table 6.1.

TABLE 6.1. Initial state of knowledge of an individual about the political status of Malaysia in 1939.

Hypotheses	Subjective probabilities expressed in percentages (%)
- Dutch colony	35
- English colony	30
- Independent State	10
- French colony	10
- Japanese colony	5
Total	90

It must be noted that, in the example of Table 6.1, the sum of SPs is not equal to 100 (as expected). This means that the student does not reject the possibility (with probability .10) that the correct answer is not among the listed ones.

To the classical researcher, our (Belgian) student will answer by the most (subjectively) probable solution: *Dutch colony*.

Such a behavior is consistent with decision theory (maximization of expected utility) and is spontaneously adopted by students. Each teacher can observe this easily.

Let us suppose, now, that our student receives successive information related to the question, since he watches a movie on "Malaysia in 1942". In this film appears a panel on which is written the word "DANGER". Since this word exists in English as in French, our student will change his subjective probabilities as in Table 6.2.

TABLE 6.2. Second state of knowledge of an individual about the political status of Malaysia in 1939.

- Dutch colony	10
- English colony	35
- Independent state	10
- French colony	30
- Japanese colony	5
Total	90

To the classical researcher, the student will answer "English colony", so that the classical researcher could conclude that this bit of information (in the movie) has drastically modified the student's cognitive state. But the classical researcher will not note that the absolute subjective probability for the answer "English colony" is almost unmodified, whereas its relative position has changed from second to first.

Let us suppose that, as the film goes on, the student sees that all the cars run on the left side of the streets. This additional information, again, changes his cognitive state as shown in Table 6.3.

TABLE 6.3. Third state of knowledge of an individual about the political status of Malaysia in 1939.

- Dutch colony	2
- English colony	90
- Independent state	2
- French colony	2
- Japanese colony	2
Total	98

The classical researcher will observe the *same external response* (English colony). He will be tempted to conclude that this bit of information is of low, if any, informative power for this student, and, as a consequence may be dropped from the movie. Such a misinterpretation derives directly from the rawness of the available data. Since the majority of the educational experiences are made using classical instructions, it is not surprising that a lot of them conclude, for example, that "there is no difference between approach A and approach B". Our belief is that in most cases differences exist, but that the experimenter was not able to observe them.

From this point of view, educationists of the early 1980's can be compared with chemists working with a coal-shovel. Since chemistry made its decisive progress through careful measurement of weights, one can easily imagine that coal-shovel chemists might hardly discover anything. Subjective probabilities will be a helpful tool in assessing the informative power of educational media. Of course, precautions must be taken to warrant validity, reliability and acuity of the measurements.

CONFIDENCE TESTING AND THE RASCH MODEL

Confidence marking should improve the estimation of ability (predictive validity would be increased) if the student is a good self-estimator. Currently, we are testing this hypothesis by a jack-knife approach, whereas a more theoretical approach in the form of a mathematical model has been undertaken by Defays (1982).

colony, English colony, Independent state, French colony, and Japanese colony. He can even do more than list or order these possible solutions: he can attribute to each a subjective probabilty of being the correct answer, as the SP researcher requested. Let us suppose that the answers are as in Table 6.1.

TABLE 6.1. Initial state of knowledge of an individual about the political status of Malaysia in 1939.

Hypotheses	Subjective probabilities expressed in percentages (%)
- Dutch colony	35
- English colony	30
- Independent State	10
- French colony	10
- Japanese colony	5
Total	90

It must be noted that, in the example of Table 6.1, the sum of SPs is not equal to 100 (as expected). This means that the student does not reject the possibility (with probability .10) that the correct answer is not among the listed ones.

To the classical researcher, our (Belgian) student will answer by the most (subjectively) probable solution: *Dutch colony*.

Such a behavior is consistent with decision theory (maximization of expected utility) and is spontaneously adopted by students. Each teacher can observe this easily.

Let us suppose, now, that our student receives successive information related to the question, since he watches a movie on "Malaysia in 1942". In this film appears a panel on which is written the word "DANGER". Since this word exists in English as in French, our student will change his subjective probabilities as in Table 6.2.

TABLE 6.2. Second state of knowledge of an individual about the political status of Malaysia in 1939.

- Dutch colony	10
- English colony	35
- Independent state	10
- French colony	30
- Japanese colony	5
Total	90

To the classical researcher, the student will answer "English colony", so that the classical researcher could conclude that this bit of information (in the movie) has drastically modified the student's cognitive state. But the classical researcher will not note that the absolute subjective probability for the answer "English colony" is almost unmodified, whereas its relative position has changed from second to first.

Let us suppose that, as the film goes on, the student sees that all the cars run on the left side of the streets. This additional information, again, changes his cognitive state as shown in Table 6.3.

TABLE 6.3. Third state of knowledge of an individual about the political status of Malaysia in 1939.

- Dutch colony	2
- English colony	90
- Independent state	2
- French colony	2
- Japanese colony	2
Total	98

The classical researcher will observe the *same external response* (English colony). He will be tempted to conclude that this bit of information is of low, if any, informative power for this student, and, as a consequence may be dropped from the movie. Such a misinterpretation derives directly from the rawness of the available data. Since the majority of the educational experiences are made using classical instructions, it is not surprising that a lot of them conclude, for example, that "there is no difference between approach A and approach B". Our belief is that in most cases differences exist, but that the experimenter was not able to observe them.

From this point of view, educationists of the early 1980's can be compared with chemists working with a coal-shovel. Since chemistry made its decisive progress through careful measurement of weights, one can easily imagine that coal-shovel chemists might hardly discover anything. Subjective probabilities will be a helpful tool in assessing the informative power of educational media. Of course, precautions must be taken to warrant validity, reliability and acuity of the measurements.

CONFIDENCE TESTING AND THE RASCH MODEL

Confidence marking should improve the estimation of ability (predictive validity would be increased) if the student is a good self-estimator. Currently, we are testing this hypothesis by a jack-knife approach, whereas a more theoretical approach in the form of a mathematical model has been undertaken by Defays (1982).

Confidence marking could be a direct experimental control of the validity of the Rasch model (see Choppin, 1980; Wright & Stone, 1979). In the Rasch model, the probability (P) of a correct answer is a function of the student's ability (A) on the one hand and of the item difficulty (D) on the other hand. In the classical approach, A and D are estimated from a matrix of responses (students - questions). The P values are computed by the famous formula:

$$P = \frac{X^{A-D}}{1 + X^{A-D}}$$

In this formula, X is a constant. Often X = e, that is 2.71828, but Choppin (1978) has pointed out that W is increasingly used (W = 1.24573) because of its interesting mathematical properties.

The "Rasch computed" P values could be directly compared to the SP values (research currently in progress).

For instance, suppose that a student has been presented a series of Rasch calibrated items (that have Rasch indices) from a given content. His successes and failures on these calibrated items enable the researcher to compute his Rasch Ability index (A value) for this content. For any Rasch calibrated item (that has a D index) presented to this student, it is now possible to compute a probability of success, P, by the formula above.

If the student is requested to provide a confidence index (or subjective probability) for each response to an item, then a correlation can be computed for each individual, between the P and SP values. The plotting of those two series of values is also of interest. If there are discrepancies between the Ps and SPs, it may be that they are not distributed over the whole range of probabilities. As will be seen below (in the Bayesian approach), predictions from a human and predictions from a formula are likely to differ systematically in some respects.

This does not mean that SPs are more trustworthy than the Rasch estimates and should be used as criteria to validate them. But obtaining two different estimates for the same probability is, of course, a situation through which the two methods (Rasch and SP) could be improved.

CONFIDENCE DEGREES AND SUBSEQUENT BEHAVIOR

It is reasonable to suspect that overt behavior is more related to the individual's beliefs than to objective measures of knowledge. The following experiment illustrates this evidence.

Lumingu (1974) presented 17 multiple choice items on word definitions to 128 thirteen year-old students. In the first stage of the experiment, the students had to answer the question (without dictionary) and to indicate their confidence degree (using the codes 0, 1, 2, and 3). In the second stage,

they were allowed to consult a dictionary (they had to note the number of the consulted pages), but were not given time enough to verify the 17 words. In the third stage, the students had to answer again (with the possibility of changing response *and* confidence degree). The analysis of the data has been undertaken by Leclercq (1975).

The general findings are not surprising. Here are seven of them:

1. It is not the objective correctness of the response that explains the use of the dictionary: it was consulted with a rate of 41% for incorrect responses (on the first stage) and of 40% for correct ones.

2. The use of the dictionary decreases with high degrees of confidence as is shown in Table 6.4.

TABLE 6.4. Rates of use of the Dictionary for each degree of confidence.

- For confidence degree 0, 47 % of use of the dictionary.
- For confidence degree 1, 45 % of use of the dictionary.
- For confidence degree 2, 40 % of use of the dictionary.
- For confidence degree 3, 33 % of use of the dictionary.

3. Consulting the dictionary helps in providing the correct answer in the post-test (83% success vs. 32% when no consulting occurred).

4. When the dictionary was *not* used, the higher the confidence degree at the pretest, the lower the changes in responses at the post-test (see Table 6.5).

5. When the dictionary was used, the higher the confidence degree at the pretest, the lower the rate of changes at the post-test (see Table 6.6).

6. The use of the dictionary improves the *average rate of success* for the various confidence degree (RSC), as can be seen in Table 6.7.

TABLE 6.5. Percentages of changes among responses accompanied by given Confidence Indices at the Pretest, when the Dictionary has not been used.

Confidence degree at the pretest	Percentages (%) of changes
0	60 %
1	58 %
2	44 %
3	37 %

TABLE 6.6. Percentages of changes among responses accompanied by given Confidence Indices at the Pretest, when the Dictionary has been used.

Confidence degree at the pretest	% of Changes
0	100 %
1	72 %
2	65 %
3	51 %

TABLE 6.7. Average Percentage Rates of Success at the Pretest and the Post-test.

RSC_s		No use of dictionary	Use of the dictionary
Confidence degree 1	PRE	26	26
	POST	21	62
Confidence degree 2	PRE	29	34
	POST	30	71
Confidence degree 3	PRE	45	43
	POST	50	87

7. The rates of use of the various confidence degrees (RUCs) show an increase in high degrees (2 and 3 summed) at the post-test when the dictionary is used (see Table 6.8).

Other interesting questions arise:
- What would be the effect of training the students in using the dictionary?

- What would be the effect of increasing the readibility of the texts (simplification of sentences, examples, drawings, ...) in the dictionary?
- What is the effect of the type of document (receipts, geography, atlas, maintenance manuals, etc.)?

TABLE 6.8. Percentage of use of High Degrees of Confidence at the Post-test when the Dictionary is used.

No use of dictionary	: 66 % of use of high degrees.
Use of dictionary	: 94 % of use of high degrees.

BAYESIAN THEORY AND THE REVISION OF PROBABILITIES

Education should be more interested in the modification of cognitive states than in a fixed cognitive state. Bayes' formula mostly used in economics, also proved to be interesting in psychology (cf. Rouanet, 1961; Edwards, 1967).

The Theorem

In this theorem, the basic data are the (subjective) probabilities the individual attributes to the various possible responses before and after receiving information.

The subjective probabilities are referred to as the *prior* probabilites and the *posterior* probabilities. The amount of information can be measured by the difference between these probabilities for the correct answer. Bayes' theorem allows us to go a step further in the analysis.

Bayes, an English clergyman, suggested in 1763, that the posterior probability of an event should be proportional to the product of the prior probability and the likelihood of this event:

$$\text{posterior } sp_i = \frac{(\text{prior } sp_i) \cdot (\text{likelihood of event } i)}{\sum_{j=1}^{n} (\text{prior } sp_j) \cdot (\text{likelihood of event } j)}$$

The likelihood of event i is the probability of information X if event i is true. It could be noted p(X | (E = true)) or p(X | E). This is sometimes called "likelihood of E given information X". The numerical value of the denominator will be referred to as DEN.

An Example

A simple example might be helpful. Consider four urns externally identical but of which the contents are different: Urn A contains 2/3 red marbles and 1/3 black ones, whereas urns B, C and D contain the inverse proportions.

An individual is informed of the contents and is presented with one of the four urns chosen at random. He has to express his (prior) subjective probability that this urn is urn A. Here the prior SPA is .25. At this point, the student is allowed further information: he is allowed to "draw" ten marbles at random out of the urn.

Suppose that he obtains 7 red and 3 black marbles. The likelihood of A is the probability of information (that is randomly pulling at least seven red marbles out of ten) if the urn is really urn A (that is containing 2/3 red marbles). Such a probability can be found in appropriate tables, for instance, the *Tables of the cumulative binomial probability distribution*, Harvard University Press, 1955 (see Table 6.9).

TABLE 6.9. Extract from Tables of the Cumulative Binomial Probability Distribution.

n	*r*	*p*=0.26	*p*=0.27	*p*=0.28	*p*=0.29	*p*=0.30	*p*=0.31	*p*=5/16	*p*=0.32	*p*=0.33	*p*=1/3
10	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	0.95076	0.95702	0.96256	0.96745	0.97175	0.97554	0.97641	0.97886	0.98177	0.98266
	2	0.77776	0.79807	0.81696	0.83449	0.85069	0.86564	0.86918	0.87938	0.89199	0.89595
	3	0.50422	0.53351	0.56217	0.59010	0.61722	0.64344	0.64985	0.66872	0.69300	0.70086
	4	0.24793	0.27258	0.29794	0.32392	0.35039	0.37724	0.38400	0.40436	0.43163	0.44074
	5	0.09035	0.10368	0.11812	0.13365	0.15027	0.16795	0.17253	0.18666	0.20635	0.21313
	6	0.02391	0.02872	0.03420	0.04039	0.04735	0.05511	0.05718	0.06371	0.07320	0.07656
	7	0.00446	0.00562	0.00700	0.00865	0.01059	0.01286	0.01349	0.01550	0.01855	0.01966
	8	0.00056	0.00074	0.00096	0.00124	0.00159	0.00202	0.00214	0.00254	0.00317	0.00340
	9	0.00004	0.00006	0.00008	0.00011	0.00014	0.00019	0.00020	0.00025	0.00033	0.00036
	10	0.00000	0.00000	0.00000	0.00000	0.00001	0.00001	0.00001	0.00001	0.00002	0.00002

The interpretation of the right column of Table 6.9 appears in Table 6.10

TABLE 6.10. Interpretation of the right-hand column of TABLE 6.9.

p of obtaining after 10 trials

3 black marbles or more (i.e. 7 red ones or less) is .70086.
2 black marbles or more (i.e. 8 red ones or less) is .89595.
1 black marbles or more (i.e. 9 red ones or less) is .98226
0 black marbles or more (i.e. 10 red ones or less) is 1.

Consequently, the "punctual probabilities" may be computed as in Table 6.11.

TABLE 6.11. Punctual probabilities.

p of obtaining exactly

7 red marbles = .70086 - .44074 = .26012
8 red marbles = .89595 - .70086 = .19509
9 red marbles = .99266 - .89595 = .08671
10 red marbles = 1 - .98266 = .01734

.55926

Note that the total (appearing in Table 6.11) equals 1 - 44074. Similarly, the p of obtaining at least 7 red marbles if the urn is B, C or D (that is, contains 1/3 red marbles) is .01966.

The probability of information (that is randomly pulling at least seven red marbles out of ten) is presented in Table 6.12.

TABLE 6.12. Probability of the Information depending on the identity of the Urn.

p(X | A) = .70086
p(X | B) = .01966
p(X | C) = .01966
p(X | D) = .01966

The numerator values of the above formula are known:

$$\text{posterior spA} = \frac{.25 \;.\; 70086}{\text{DEN}}$$

To compute the denominator, it must be remembered that post spA + post spB + post spC + post spD must be equal to 1.

So we may write that:

posterior psA = .25 x .70086 / DEN
posterior psB = .25 x .01966 / DEN
posterior psC = .25 x .01966 / DEN
posterior psD = .25 x .01966 / DEN

1 = 1

Consequently, 1 = (.25 x .70096/DEN) + 3(.25 x .01966/DEN) - and DEN = .18996. It follows that the posterior spA = .92337, and that the posterior spB (or C or D) = .02587. Note that the sum of the four posterior SPs is close to 1.

The Use of the Theorem

Such a theorem has been used for medical diagnosis. Lindley (1971, p.104) notes that even before the patient has entered the consulting room, the physician has *prior* probabilities on what could be wrong with his visitor. Indeed, a lot of Europeans suffer, in Europe, from a sore throat but very few are attacked by beriberi!

The physician gathers information by questioning the patient, by examining him and by making tests. The doctor has now to combine information and prior probabilities. If the set of symptoms X is rarely associated with a given disease, its likelihood will be low and consequently, the product (on the numerator) will be low too. Even if he has never heard of Bayes' theorem, a doctor processes his information largely according to it. Interactive help to medical diagnosis might work as follows. Prior SPs of various diseases (and likelihoods of the diseases given various bits of information) would be stored in the computer. What is missing is the nature of information X. This is precisely, what would be entered by the doctor into the computer. The obvious advantage of such an approach is that the relevant computations are quickly and correctly done and that the whole medical experience, and not only this doctor's, would be available.

From the theorem, it appears that one should not attribute zero prior probability values to uncertain events. In this case, the posterior probabilities would also be zero, whatever the information, which would be absurd. A prior probability value of one is equally dangerous (because the other events have zero prior probabilities).

Lack of complete certainty does not imply that people are reduced to inaction. We continue to drive cars, fly on jets, cross streets whereas we know the probability of being killed in these actions is not zero ... and, in some cases, we remain alive precisely because we are aware of the dangers.

Bayes' theorem raises other problems related to education; for instance, the credibility of estimators or experts or teachers, that is, the sureness of the "source of information" (see Lindley, 1971). The theory states how we should learn but not how we actually do. There, it is normative, not descriptive. What is the reality? A tempting approach will be found in the next section.

EXPERIMENT TO ASSESS SUBJECTS INFORMATION PROCESSING CAPACITIES

Classical Findings

The crucial problem is: do human beings revise their probabilities according to Bayes' theorem? The answer is definitely no. Many researchers have observed what Edwards, Lindman and Phillips (1965) noted as follows:

> Men are incapable of extracting all of the certainty from information Bayes' theorem indicates is in that information. To put it another way, men are conservative information processors... Whatever the merits or demerits of a built-in tendency to conservatism in information processing in daily life, such a tendency is clearly a hindrance to human effectiveness in information processing systems.... Consequently, the finding of human conservatism raises some problems for the design of man-machine systems intended to perform information processing in a more or less optimal way.

Human conservatism appears clearly in Figure 6.1 (from Edwards, 1967).

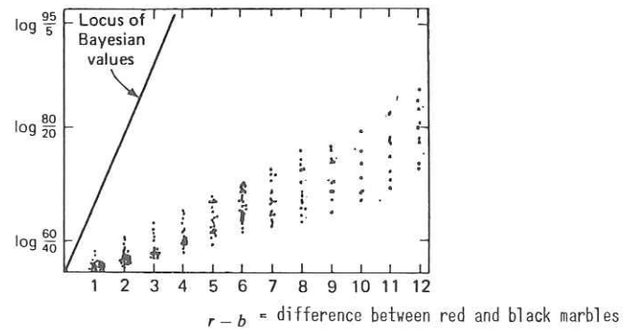


Fig. 6.1. Probability estimates of a subject compared with Bayesian estimates.

Explanations of such conservatism are numerous. Coombs, Dawes and Tversky (1970, p. 148) insist on inconsistency: "when faced with the same alternatives, under seemingly identical conditions, people do not always make the same choice." Rouanet (1961) has suggested another explanation of this "under Bayesian" efficiency: when there is a discrepancy between prior probabilities and information, the individual has a tendency to ignore one of the two criteria. In one of his experiments, the subjects neglected the prior

knowledge of conceptual nature whereas they took into account only the information of a more perceptual nature. Moreover, contradictions between prior probabilities and information provoke cognitive *dissonance* and the mechanism of reduction of dissonance (see Festinger, 1964).

An Experiment Close to the Educational Type

As Edwards (1967) has noted, the problem with the Bayesian theorem is that in every-day life situations the necessary data do not exist or are not objectively attributable. The solution consists of asking judges or experts to assess (subjectively) these probabilities. Obviously, there is a large gap between urns filled with marbles and common cognitive content encountered in educational situations.

In order to explore the difficulties linked with educational situations, a very simple "revision guessing game" (RGGame) has been developed. In this game, the subjects are presented with truncated sentences and have to provide the first hidden letter. As compared with the CGGame earlier, six (plausible) alternatives are presented there, the correct answer being one of them.

First step. The subject must provide a subjective probability (sp) for each of the six letters (with $\sum sp = 1$). The whole game contains five sentences and the tariffs are the same as for the CGGame. An example is presented in Table 6.13.

TABLE 6.13. Example of the First Step of the RGGame. The Subject has to distribute 100 % probabilities on six alternatives. The complete sentence is "THERE IS NO REVERSE ON A MOTORCYCLE", the correct answer is V.

```
THERE IS NO RE
the following letter is A    P    V    L    S    T
                     sp : ..   ..   ..   ..   ..   ..
```

Second step. The subject is told that now a letter will be chosen randomly from a printed table, and that he will be requested to assess the probability of this random letter following the unknown letter. The subject is advised that the randomness of the procedure will produce awkward combinations. Actually, the so called random letter is not at all randomly chosen. It is the letter that really follows the unknown letter (in the example, this "random letter" is E). This second step is undertaken for the five sentences. An example of this is presented in Table 6.14.

TABLE 6.14. Example of the second Step of the RGGame.

```
        THERE IS NO RE
                     ↑
....    = probability of E following A if A is the unknown letter ?
....    = probability of E following P if P is the unknown letter ?
....    = probability of E following V if V is the unknown letter ?
Etc.
```

Third step. The subjects are told that the "random letter" is, in reality, the following letter; this is the *information* of the Bayesian approach. The subjects have to provide posterior probabilities to each of the six letters knowing the following one (for instance, letter A now has a low probability).

Results

During the experiment, it appeared that subjects were not able to assess *conditional* probabilities and, in fact, assessed *joint* probabilities. For instance, they did not assess the probability of the random letter (here E) following P *given that* P is the unknown letter. In fact, they assessed the probabiltiy of the *couple* of letters PE.

A radical change had to be introduced in the game, by inverting steps 1 and 2 in order to have the conditional probability likelihoods assessed before the prior probabilities. The resulting instructions were more complicated than the original ones. Meanwhile, the number of alternatives had been reduced from six to three whereas the number of sentences increased from five to ten.

This second version allowed us to gather 30 complete sets of data for each of the six individuals. Six correlations were computed (on 30 reponses each) between posterior probabilities from the human processor and posterior probabilities from Bayes' theorem. Those correlations are presented in Table 6.15.

TABLE 6.15. Correlations between two kinds of Posterior Probabilities : the subjective ones and the ones computed with Bayes' theorem.

Subject 1 :	.85
Subject 2 :	.67
Subject 3 :	.81
Subject 4 :	.95
Subject 5 :	.47
Subject 6 :	.84

The game produced a majority of "close to zero" and "close to one" posterior probabilities. In these cases, man and Bayesian theorem results are very similar.

The game could not prevent the subjects from introducing quite new hypotheses (provided by insight) between the prior and posterior assessments. The proximity of the two kinds of estimation of "intermediate" posterior probabilities (from .20 to .80) is deceivingly low.

For the lowest correlation (.47), it appears that the discrepancies are due to two questions.

In the scatter diagram of Figure 6.2, the three arrows indicate three couples of values produced by two questions.

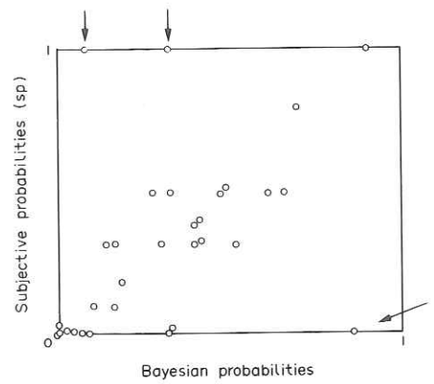


Fig. 6.2. Graphic representation of the thirty pairs of values (three for each of ten questions) for Subject 5 in the experiment.

The correlation computed on the eight remaining questions is .91.

CONCLUSION

It appears again that the quality of data relies on the quality of the instruments. The complexity of cognitive problems involves the need for sophisticated tools. That is the price psychology and education have to pay to deepen the understanding of their field, in the same way as chemistry and biology did previously.

GENERAL CONCLUSIONS

Research on subjective probabilities in education raises theoretical and methodological problems similar to those that rose in psychology when psychophysics developed. Strict definitions had to be found for "stimulus" and for "response". In the same way, concepts like "knowledge", "doubt", "uncertainty", "confidence", "information", begin to have an operational meaning in a subjective probability framework.

We have argued that subjective probability could enable us to assess partial knowledge with valid, reliable, sensitive and convenient methods and techniques. Specific principles must be respected in this kind of approach, and clear distinctions must be made between measures, payoffs and degrees. When this is done, promising perspectives appear concerning the study of information processing by the human, a central concern for educators. Steps further have already been made in quantifying the amount of transmitted information (Van Naerssen, 1965; Dirkwazer, 1978).

Most of the research reported in the previous chapters are only indicative since new hypotheses had to be tested, instruments had to be tried, indices had to be explored. As a result, data are not representative of populations.

As well as individual differences, the whole context (cultural, sociological, economical, political) has an important impact on the subjective probabilities and risk taking behavior. For instance, the tendency to guess at multiple choice varies from country to country. No macro analysis of the like has been undertaken in the present study although it would be very interesting to deepen such points as the differential consequences of the same educative actions in different human contexts. The same can be said for differences as age, school or intellectual level, sex, tiredness, competency in the field, various kinds of training, and, first of all, personality. Correlations between subjective probability and response time would be of interest too.

Other research could explore inter-item subjective consistency, as well as inter-behavior consistency (for example, relations between opinions, beliefs, behavioral intentions, and behavior).

In another domain, classical experiments in psychophysics could be revised, including subjective probabilities as methods of responses. The same could be said for stochastic models of learning (Rouanet, 1965).

The present work does not provide conclusions, but we would be pleased if it convinced readers that there is an important domain to explore, that trustworthy methods exist and that there are more reasons for deepening our knowledge than for remaining on the surface of these phenomena.

REFERENCES

Adams, J.K., & Adams, P.A., Realism of confidence judgments. *Psychological Review*, 1961, 68, 33-45.

Ahlgren, A., *Confidence on achievement tests and the prediction of retention*. Unpublished doctoral dissertation, Harvard University, 1967.

Allais, M., Le comportement de l'hommme rationnel devant le risque: critique des postulats et axiomes de l'école américane, *Econometrika*, 1953a, 21, 503-546.

Allais, M., La psychologie de l'homme rationnel devant le risque: La théorie et l'expérience, *Journal of Social Statistics*, 1953b, 94, 43-73.

Atkinson, J.W., *An Introduction to Motivation*, Princeton: Van Nostrand, 1964.

Atkinson, J.W., & Litwin, G.H., Achievement motive and test anxiety conceived as motive to approach success and motive to avoid failure, *Journal of Abnormal Social Psychology*, 1960, 60, 52-63.

Attneave, F. *Applicatoin of information theory to psychology*, New York: Holt Rinehart and Winston, 1959.

Baker, J.D., The uncertain student and the understanding computer, *La Recherche en enseignement programmé, Tendances actuelles*. Paris: Dunod, 303-319, 1969.

Bartholomé, M., & Houziaux, M.O. *SIAM-DOCEO II, Instruction manual*, 1979.

Bayes, T., An essay toward solving a problem in the doctrine of chance, *Philosophical Transactions of the Royal Society*, London, 1763.

Beaujot, A., Didelez, M., Fontaine, O, & Leclercq, D., Etude d'une nouvelle technique d'évitement sans signal avertisseur chez le rat, *Psychologica Belgica*, 1966, VII.

Beenen, F., Psychiatric diagnosis and subjective probabilities, *Acta Psychologica*, 1970, 34.

Bernouilli, D. Exposition of a new theory on the measurement of risk (English translation of "Specimen theorial novea de mensura sortis", Commentarii academiae scientiarum imperialis Petropolitanae, 1738, 5, pp. 175-192), by Louise Sommer), *Economica*, 22, 23-26.

Brier, G.W., Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, 1950, 75, 1-3.

Brownless, V.T. & Keats, J.A. A retest method fo studying partial knowledge, *Psychometrika*, 1958, 23, 1, 67-73.

Bujas, Z., Koviacic M., & Rohacek, A., Psychological function based on confidence rating, *Acta Instituti Psychologici Universitatis Zagrabiensis*, 1975, 74.

Chauvin, R, Paradoxe dans les résultats du conditionnement, *Journal de Pschychologie Normale et Pathologique*, 1967, 129-141.

Chernoff, H., Rational selection fo decision functions, *Econometrica*, 1954, 422-443.

Chernoff & Moses, *Elementary Decision Theory*, New York: Wiley, 1954.

Chernoff, H., The scoring of multiple choice questionnaires, *Annals of Mathematical Statistics*, 1962, *33*, 375-393.

Cherry, E.C., On the validity of applying communication theory to experimental psychology, *British Journal of Psychology*, 1957, *48*, 176-188.

Choppin, B., *An IEA Study of guessing. A Proposal*. Stockholm, International Association for the Evaluation of Educational Achievement. Unpublished memorandum, IEA/TR/9.

Choppin, B., *The correction for guessing on objective tests*, Stockholm, International Association for the Evaluation of Educational Achievement, 1974.

Choppin, B., Guessing the answer on objective tests, *British Journal of Educational Psychology*, 1975, *45*, 206-213.

Choppin, B., Recent developments in item banking. A review. Montreux, Second International Symposium on Educational Testing, 1975.

Choppin, B., Item banking and the monitoring of achievement. Research in progress, NFER Series, April, 1978.

Clark, R.A., Teevan, R., Ricciuti, H.N., Hope of success and fear of failure as aspects of need for achievement, *Journal of Abnormal Social Psychology*, 1956, *53*, 182-186.

Cooke, W.E, Forecasts and verifications in Western Australia, *Monthly Weather Review*, 1906a, *34*, 23-24.

Cooke, W.E., Weighting forecasts, *Monthly Weather Review*, 1906b, *34*, 274-275.

Coombs, C.H., Psychological scaling without a unit of measurement, *Psychological Review*, 1950, *57*, 145-158.

Coombs, C.H., On the use of objective examinations, *Educational and Psychological Measurement*, 1953, *13*, 108-130.

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

Coombs, C.H., & Pruitt, Components of risk in decision making Probability and variance Preferences, *Journal of Experimental Psychology*, 1950, 265-277.

Coombs, C.H., A Theory of data, *Psychological Review*, 1960, *67*, 143-159.

Coombs, C.H., Greenberg, M.G., & Zinnes, J.A., A double law of comparative judgment for the analysis of preferential choice and similarities data, *Psychometrika*, 1961, *26*, 165-171.

Coombs, C.H., *A Theory of data*, New York: Wiley, 1964.

Coombs, C.H., Dawes, R.M., Tversky, A., *Mathematical Psychology*, Englewood Cliffs, NJ: Prentice Hall, 1970.

Coombs, C.H., & Bowen, J.N., A test of VE theories of risk and the effect of the central limit theorem, *Acta Psychologica*, 1971, 35, 15-28.

Crawford, W.R., & Lewy, A., *A rapid and efficient method for scoring and analyzing complex multiple choice examinations*, Chicago: National Council on Measurement in Education, 1965.

Cronbach, L.J., Further evidence on response sets and test design, *Psychological Measurement*, 1950, *10*, 3-31.

Cronbach, L.J., & Meehl, P.E., Construct validity in psychological tests, *Psychological Bulletin*, 1955, *52*, 281-302.

Davis, F.B., Use of correction for chance success in test scoring, *Journal of Educational Research*, 1959, *52*, 179-180.

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

Davis, F.B., *Educational Measurements and their Interpretations*, Belmont, CA: Woodsworth, 1964.

De Finetti, B., La prévision: ses lois logiques, ses sources subjectives, *Annales de l'Institut Henri Poincaré*, 1937, 7.

De Finetti, B., Dans quel sens la théorie de la décision est-elle et doit-elle être normative? In F.N.R.S. (Ed.) *La Décision*, Paris: FNRS, 1959.

De Finetti, B., Does it make sense to speak of good probability appraisers? In Good, I.J. (Ed.), *The Scientist Speculates*, New York: Basic Books, 1962, 357-364.

De Finetti, B., La décision et les probabilités, *Revue des Mathématiques pures et appliquées*, Bucarest, 1963, 405-413.

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.

D'Hainaut, L, Une méthode de compensation statistique des choix heureux par ignorance dans les questions fermées d'épreuves d'acquisition, *Les sciences de l'éducation*, 1974, *7*, 1, 57-83.

Diamond, J. & Evans, W., The correction for guessing, *Review of Educational Research*, 1973, *43*, 2.

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

Ebel, R.L., *Measuring Educational Achievement*, Englewood Cliffs, NJ: Prentice Hall, 1965a.

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

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

Ebel, R.L., Expected reliability as a function of choices per item, *Educational and Psychological Measurement*, 1969, *29*, 565-570.

Ebert, R., Sequential decision making: An aggregate scheduling methodolgy, *Psychometrica*, 1971, *36*.

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

Edgington, E.S., Soring formulas that correct for guessing, *Journal of Experimental Education*, 1965, *33*, 345-346.

Edwards, W., Probability preferences in gambling, *The American Journal of Psychology*, 1953, *66*, 349-364.

Edwards, W., Variance preferences in gambling, *American Journal of Psychology*, 1954, *67*, 441-452.

Edwards, W., Probabiltiy preferences among bets with differing expected values, *American Journal of Psychology*, 1954, *67*, 56-57.

Edwards, W., The reliability of probability preferences, *American Journal of Psychology*, 1954, *67*, 68-95.

Edwards, W., The theory of decision making, *Psychological Bulletin*, 1954, *51*, 380-417.

Edwards, W., Methods for computing uncertainties, *American Journal of Psychology*, 1954, *67*, 164-170.

Edwards, W., The prediction of decisions among bets, *Journal of Experimental Psychology*, 1955, *59*, 201-214.

Edwards, W., Measurement of utility and subjective probability, in H. Gulliksen & Messick, (Eds.) *Psychological Scaling: Theory and Applications*, New York: Wiley, 1960.

Edwards, W., Probability learning in 1000 trials, *Journal of Experimental Psychology*, 1961, *62*, 4, 385-394.

Edwards, W., Behavioral Decision theory, *Annual Review of Psychology*, 1961, *12*, 473-498.

Edwards, W., Subjective probabilities inferred from decisions, *Psychological Review*, 1962, *69*, 109-135.

Edwards, W., Utility, subjective probability, their interaction and variance preferences, *J. Conf. Res*, 1962, *6*, 42-51.

Edwards, W., Probabilistic information processing by men and man-machine systems. In *La simulation du comportement humain*, Paris: Dunod, 1967, pp. 187.

Edwards, W., Lindman, H., & Phillips, L.D., Emerging technologies for making decisions, in *New Directions in psychology*, Vol. II, New York: Holt, Rinehart and Winston, 1965, 261-325.

Epstein, E.S., A scoring system for probability forecasts of ranked categories, *Journal of Applied Meteorology*, 1969, *8*, 985-987.

Fabre, J.M., Docimologie et évaluation par questionnaires: étude du jugement multiple et de l'autopondération, Thèse de doctorat de 3e cycle en psychologie, Université de Provence, 1977.

Festinger, L, *Conflict, Decision and dissonance*, Stanford: Stanford University Press, 1964.

Fischer, G., Tailored testing on the basis of the Rasch model. Paper presented at the 3rd International Symposium on Educational Testing, Leyden, 1977.

Greenberg, M., J scale models for preference behavior, *Psychometrika*, 1963, *28*, 3, 265-271.

Hambleton, R.K., Roberts, D.M., & 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 Measurmeent*, 1970, *7*, 75-82.

Hamilton, C.H., Bias and error in multiple-choice tests, *Psychometrika*, 1950, *15*, 151-168.

Hammerton, M., The guessing correction in vocabulary tests, *British Journal of Educational Psycholgoy*, 1965, *35*, 249-251.

Hancock, J.G., & Teevan, R.C., Fear of failure and risk-taking behavior, *J. Pers*. 1964, *32*, 200-209.

Hardy, J.L, *Approche expérimentale du comportement d'estimation et de sa mesure*, Unpublished graduate dissertation, University of Liège, 1980.

Hardy, J.L., Using computer-based feedback to improve estimation ability. IFIP, NCCE, Lausanne, 1981.

Henmon, V.A.C., The relation of the time of a judgment to its accuracy, *Psychological Review*, 1911, *18*, 186-201.

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.

Hollingworth, R.L., Experimental studies in judgment, *Archives of Psychology*, 1913, *29*, 1-119.

Hopkins, K.D., Extrinsic reliability, estimating and attenuating variance from response styles, chance and other irrelevant sources, *Educational and Psychological Measurement*, 1964, *24*, 271-281.

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

Horst, P., The difficulty of a multiple-choice test item, *Journal of Educational Psychology*, 1933, *24*, 229-232.

Houziaux, M.O., Les fonctions didactiques de DOCEO. In *Actes du XII Colloque de l'association internationale de pédagogie expérimentale de langue francaise*, University of Caen, 1965, 47-71.

Houziaux, M.O., *Vers l'enseignement assisté par ordinateur*, Paris: Presses Universitaires Francaises, 1972.

Houziaux, M.O., Godart, C., Lavigne, M., Bartholome, M., Luyckx, A., & Lefebvre, P., Une expérience d'enseignement assisté par ordinateur chez des patients diabetiques insulinodépendants, *Scientia Paedagogica Experimentalis*, 1978, *15*, 215-250.

Hurwicz, L. *Optimality Criteria for Decision Making under Ignorance*. 1951. Technical report no. 70, Cowles commission discussion paper, Statistics.

Isaacson, R.L., Relation between achievement, test anxiety and curricular choices, *Journal of Abnormal Social Psychology*, 1964, *68*, 447-452.

Irwin, W.S., & Smith, W.A., Value, cost and information as determiners of decision, *Journal of Experimental Psychology*, 1957, *54*, 229-232.

Jacobs, S.S., *An empirical investigation of the relatinship between selected aspects of peronality and confidence-weighting behaviors*, Doctoral dissertation, University of Maryland, Univeristy Micro-films, 68, 16, 676, 1968.

Jacobs, S.S., Correlates of unwarranted confidence in response to objective test items, *Journal of Educational Measurement*, 1971, *8*, 1.

Jungermann, A., & Dezeeuw, G., (Eds.), *Decision making and change in human affairs. Proceedings of the fifth research conference on subjective probabitiy, utility and decision making*, Darmstadt: Reidel, 1977.

Kido, J.B., The utilization of subjective probabilities in production planning, *Acta Psychologica*, 1970, *34*, 338-347.

Koehler, R.A., A comparison of the validities of conventional choice testing and various confidence marking procedures, *Journal of Educational Measurement*, 1971, *8*, 4.

Kuder, G.F., Identifying the faker, *Personal Psychology*, 1950, *3*, 155-167.

Leclercq, D., Sequential adaptive tailored testing and confidence marking, in Vanderkamp, Langerak, De Gruyter, *Psychometrics for Educational Debates: Proceedings of the 3rd International Symposium on Education Testing*, 1977, p. 306.

Leclercq, D., Concepts, procedures and coefficients to be used with confidence marking. Paper presented at the 8th European Mathematical Psychology meeting, Saarbrucken, 1977b.

Leclercq, D., L Matrices or the computations of consequences for confidence marking procedures in educational settings; rationale, algorithm and FORTRAN program. Paper presented at the 6th Research Conference on Subjective Probability, Utility and Decision-Making, Warsaw, 1977c.

Leclercq, D., Test-retest replication and spontaneous acuity of subjective probabilities; results from a guessing game. Paper presented at the 7th research conference on subjective probability, utility and decision-making, Gothenburg, 1979.

Leclercq, D., Un module d'auto-évaluation ou Comment impliquer l'etudiant dans la régulation de ses apprentissages, *Education*, 1978a, No. 165, 59-73.

Leclercq, D., L'Auto-évaluation des compétences dans le domaine cognitif, *Revue, 13e annee*, 1978b, No. 2, February, pp. 3-20.

Leclercq, D., Computerised tailored testing: Structured and calibrated item banks for summative and formative evaluation. *European Journal of Education*, 1980, *15*, 3.

Lefebvre, P., & Houziaux, M.O., Anamnèse assistée par ordinateur en diabétologie. Résultats préliminaires, *Revue Médicale de Liège*, 1969, *24*, 803-809.

Lewy, A, & McGuire, C., A study of alternative approaches in estimating the reliability of conventional tests. Paper presented at the AERA annual meeting, Chicago, 1966.

Lieblich, A., The effect of Stress and the motivation to succeed on test risk, *Journal of Personality*, 1968, *36*, 608-615.

Lichtenstein, S., Fischhoff, B., & Phillips, L., Calibration of probabilities: The state of the art. In Jungermann & DeZeeuw (Eds), 1977.

Linder, D., Wortman, C., & Brehm, J.W., Temporal changes in predecision preferences among choice alternatives, *Journal of Peronality and Social Psychology*, 1971, *19*, 282-284.

Lindley, D.V., *Introduction to probabiltiy and statistics from a Bayesian viewpoing, Part 1: Probability*, London: Cambridge University Press, 1969.

Lindley, D.V., *Introduction to Probability... Part 2: Inference*, London: Cambridge University Press, 1970.

Lindley, D.V., *Making decisions*, London: Wiley, 1971.

Littig, L.W. *The Effect of Motivation on Probability Preference and Subjective Personality*, University of Michigan, 1959,

Little, E., & Creaser, J., Uncertain responses on multiple-choice examinations, *Psychogical Reports*, 1966, *18*, 801-802.

Lord, F.M., Formula scoring and validity, *Educational and Psycholgoical Measurement*, 1963, *23*, 663-672.

Lord, F.M., The effect of random guessing on test validity, *Educational and Psychological Measurement*, 1964, *24*, 745-747.

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

Lord, F.M., Some test theory for tailored testing, in Holtsman, W.H., (Ed.) 1970.

Lord, F.M., The self-scoring flexilevel test, *E.T.S. Research Bulletin*, 1970, *70*, 43.

Lovie, A.D., & Davies, D.M., The effect of rate of revision and initial revision on the perception of another's age, *Acta Pscyhologica*, 1970, *34*, 322-327.

Luce, R.D., *Individual Choice Behavior*, New York: Wiley, 1959.

Luce, R.D., Raiffa, H., *Games and Decision*, New York: Wiley, 1966.

Lumingu, B., *Etude préalable a la construction d'un test diagnostique sur la consultation du dictionnaire*, Unpublished thesis, University of Liège, 1974.

Lyerly, S.B., A Note for correcting for chance success in objective tests, *Psychometrika*, 1951, *16*, 21-30.

Manz, W., Experiments on probabilistic information processing, *Acta Psychologica*, 1970, *34*, 184-200.

Martin, J.J., *Bayesian Decision Problems and Markov Chains*, New York: Wiley 1967.

Massengill, H.E., & Shuford, E.H., *What Pupils and Teachers Should Know About Guessing*, Lexington, MA: Shuford-Massengill Corp., 1967.

Massengill, H.E., & Shuford, E.H., *A Report on the Effect of Degree of Confidence in Student Teaching*, U.S. Air Force, Office of Scientific Research, 1968.

Medley, D.M., The effects of heterogeneity of content and guessing on the accuracy of scores in multiple-choice tests, *Americal Educational Research Journal*, 1966, *3*, 27-33.

Mellenbergh, G.J., Nieuwe Ervaringen met een Zekerheidsaanduiding, *Ned. T. Psychol.*, 1967, *22*, 168-181.

Meuwese, W., Barendregt, J.T., & Vastenhout, J., Een onderzoek naar de relatie tussen de juistheid van oordelen en het begeleidend gevoel van zekerheid, *Ned. T. Psychol.*, 1960, *15*, 529-541.

McClelland, *Studies in Motivation*, New York: Appleton, 1955.

McNeel, S.P. & Messick, D.M., A Bayesian analysis of subjective probabilities of interpersonal relationships, *Acta Psychologica* 1970, *34*, 311-321.

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

Miller, G.A., The magical number seven, plus or minus two, *Psychological Review*, 1956, *63*, 81-97.

Murphy, A.H., A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation, *J. Applied Meteorology*, 1966, *5*, 534-537.

Murphy, A.H., *The evaluation of probabilistic predictions in meteorology*. Unpublished doctoral dissertation, University of Michigan, 1969a.

Murphy, A.H., Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratio is incomplete, *Journal of Applied Meteorology*, 1969b, *8*, 863-873.

Murphy, A.H., The ranked probability score and the probability score: A comparison, *Monthly Weather Review*, 1970, *98*. Murphy, A.H. On expected-utility measures in cost-loss ratio decision situations, *J. Appl. Meteorol.*, 1969c, *8*, 989-991.

Murphy, A.H. Scalar and vector partitions of the probability score (Part 1). Two-state situation. *Journal of Applied Meteorology*, 1972, *11*, 273-282.

Murphy, A.H. A new vector partition of the probability score. *Journal of Applied Meteorology*, 1973, *12*, 595-600.

Murphy, A.H. A sample skill score for probability forecasts. *Monthly Weather Review*, 1974, *102*, 48-55.

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

Myers, A.E. Risk taking and academic success and their relation to an objective measure of achievement motivation. *Educational Psychology Measurement*, 1965, *25*, 355-363.

Oskamp, S., The relationship of clinical experience and training methods to several criteria of clinical prediction, *Psychological Monographs*, 1962, *76*.

Pitz, G.F. Subjective probability distributions for imperfectly known quantities. In L.W. Gregg (Ed.), *Knowledge and Cognition*. New York: Wiley, 1974.

Raiffa, H. *Decision Analysis, Introductory Lectures on Choice under Uncertainty*, New York, Addison-Wesley, 1970.

Richelle, M. Malentendus sur les apports du conditionnement, *Rev. Comp. Animal*, 1970, *4*, 1, 22-31.

Rippey, R.M. A Fortran Program for scoring and analyzing probabilistic tests. *Behavioral Science*, 1968, *13*, 424.

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*, 3.

Rouanet, H. Etudes de decisions experimentales et calcul de probabilites. In *La decision*, 1961, 33-43, Paris, C.N.R.S.

Ruch, G.M. & Stoddard, G.D. Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology*, 1925, *16*, 89-103.

Ruch, G.M. & DeGraaff, M.H. Corrections for chance and guess vs. do not guess. Instructions in multiple-choice tests. *Journal of Educational Psychology*, 1926, *17*, 368-375.

Sandbergen, S. Test strategie/test strategy. *Ned. T. Psychol.*, 1968, *23*, 16-38.

Sandbergen, S. *Meningen van Studenten over Zekerheidscoring/Students Opinions about Confidence Marking*, R.I.T.P. memorandum (unpublished), 1972.

Sandbergen, S. Guessing and confidence in testing educational achievement. In Choppin, B. (*A/106 IEA Memorandum*). 1972.

Savage, L.J. *The Foundations of Statistics*, New York, Wiley, 1951.

Savage, J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 1971, *66*, 336; 783-801.

Schaefer, R.E. & Borcherding, K. The assessment of subjective probability distribution: A training experiment. *Acta Psychologica*, 1973, *37*, 117-129.

Schum, D.A., Goldstein, I.L., Howell, W.C. & Southard, J.F. Subjective probability under several cost payoff arrangements. *Org. Behav. Hum. Perform.*, 1967, *2*, 84-104.

Shannon, C.E. A mathematical theory of communication. *Bell System Technical Journal*, 1948, *27*.

Shannon, C.E. & Weaver, W. *The Mathematical Theory of Communication*, University of Illinois Press, 1949

Shannon, C.E. Prediction and entropy of printed English, *Bell Syst. Techn. Journal*, 1951, *30*, 50-64.

Sherrifs, A.C. & Boomer, D.S. Who is penalized by the penalty for guessing, *Journal of Educational Psychology*, 1954, *45*, 81-90.

Shuford, E.H. *How to Shorten a Test and Increase its Reliability and Validity*. Lexington, Shuford-Massengill Corporation, Technical Report SCM R-8, 1967.

Shuford, E. Systems of confidence weighting, theory, and practice, Los Angeles, Annual Meeting of the American Educational Research Association, 1969.

Shuford, E., Albert, A., & Massengill, N.E. Admissable probability measurement procedures, *Psychometrika*, 1966, *31*, 125-145.

Shuford, E. & Brown, T.A. Elicitation of personal probabilities and their assessment. *Instructional Science*, 1975, *4*, 137-188.

Sidman, M. Avoidance conditioning with brief shock and no exteroceptive warning signal. *Science*, 1953.

Siegel, S., Siegel, A.E. & McMichael, A.J. *Choice, Strategy and Utility*, New York, McGraw-Hill, 1961.

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

Slakter, M.J. The penalty for not guessing, *Journal of Educational Measurement*, 1968, *5*, 141-144.

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

Slovic, P., Lichtenstein, S. & Edwards, W. Boredom induced changes in preferences among bets, *American Journal of Psychology*, 1968, *78*, 208-217.

Slovic, P. & Lichtenstein, S. Relative importance of probabilities and payoffs in risk taking, *Journal of Exper. Psych.*, 1968, *78*,

Smith, C.P. Relationship between achievement-related motives and intelligence, performance level, and persistence, *J. Abnorm. Soc. Psych.*, 1964, *68*, 523-533.

Smith, R.B. An empirical investigation of complexity and process in multiple-choice items, *Journal of Educational Measurement*, 1970, *7*, 1.

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

Solomon, H. *Studies in Item Analysis and Prediction*, Stanford University Press, 1961.

Stanley, J.C. & Wang, M.C. *Differential Weighting. A Survey of Methods and Empirical Studies*. New York, College Entrance Exam. Board, 1968.

Stanley, J.C. & Wang, M.D. Weighting test items and test-item opinions. *Educ. and Psych. Measurement*, 1970, *30*, 21-35.

Stevens, S.S. *Handbook of Experimental Psychology*. New York, Wiley, 1951.

Stevens, S.S. On the psychophysical law. *Psychological Review*, 1957, *64*, 153-181.

Stevens, S.S. Measurement, psychophysics and utility. In C.W. Churchman & P. Ratoosh (Eds.), *Measurement, Definitions and Theories*, New York, Wiley, 1959.

Stevens, S.S. The surprising simplicity of sensory metrics. *Amer. Psychol.*, 1962, *17*, 29-39.

Swineford, F. The measurement of a personality trait. *Journ. of Educ. Psych.*, 1938, *29*, 289-292.

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

Swineford, F. & Miller, P.M. Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test. *Journal of Educational Psychology*, 1953, *44*, 129-133.

Tables of the Cumulative Binomial Probability Distribution. Cambridge, Mass., Harvard Univ. Press, 1955.

Thorndike, R.L. & Hagen, E. *Measurement and Evaluation in Psychology and Education*. New York, Wiley, 3e ed., 1969.

Thorndike, R.L. *Educational Measurement*. Amer. Council on Education, 2e ed., 1971.

Thrall, R.M., Coombs, C.H., & Davies, R.L. *Decision Processes*. New York, Wiley, 1954.

Tiberghien, G. Etude de la certitude du rappel au cours d'un apprentissage verbal. *Année psychol.*, 1968, *18*, 32-39.

Torgerson, W. *Theory and Methods of Scaling*. New York, Wiley, 1967.

Trow, W.C. The psychology of confidence, an experimental inquiry. *Archives of Psychology*, 1923, *67*, 1-47.

Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 1974, *185*, 1124-1131.

Van Naerssen, R.F. A scale for the measurement of subjective probability. *Acta Psychologica*, 1962, *20*, 2, 159-166.

Van Naerssen, R.F. & Van Beaumont, R. Ervaringen met een Zekerheidsaanduiding bij objektieve Tentamens. *Ned. T. Psychol.*, 1965, *20*, 308-315

Van Naerssen, R.D., Sandbergen, S. & Bruynis, E. Is de Utiliteitscurve van Examenscores een Ogief? *Ned. T. Psychol.*, 1966, *21*, 6, 358-363.

Von Neumann, J. & Morgenstern, D. *Theory of Games and Economic Behavior*. Princeton Univ. Press, 1947.

Votaw, D.F. The effect of Do-Not-Guess directions upon the validity of true-false or multiple-choice tests. *Journ. of Educ. Psychol.*, 1936, *28*, 698-703.

Waters, C.W. & Waters, L.K. Validity and likeability ratings for three scoring instructions for a multiple-choice vocabulary test. *Test, Educ. and Psych. Measur.*, 1971, *31*, 935-938.

Wiley, L.N. & Trimble, D.C. The ordinary objective test as a possible criterion of certain personality traits, *School and Society*, 1936, *43*, 446-448.

Williamson, J. Assessing clinical judgment. *J. of Medical Educ.*, 1964, *39*, 893.

Winkler, R.L. The quantification of judgment: Some methodological suggestions. *J. Amer. Statist. Ass.*, 1967, *62*, 1105-1120.

Winkler, R.L. Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Ass.*, 1969, *64*, 1073-1078.

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

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

Wood, R. Multiple choice: A state of the art report. In B. Choppin & T.N. Postlethwaite, *Evaluation in Education: International Progress*. Pergamon, 1977.

Wright B.D. & Stone M.H. *Best Test Design*. Mesa Press, Chicago, 1979.

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

Aims and Scope

This series of refereed monographs in educational evaluation is designed to make available to a wide audience: state of the art reviews; new evaluation methodologies, and selected specific reports on important topics.

The series includes contributions from throughout the world in order to highlight new concepts, methods and approaches being undertaken in a particular country or region but not known elsewhere. Different countries (developed and developing) can learn from each other about the way in which evaluation is used as a basis for improving education. The way in which evaluation projects are planned, executed and the results translated into action in different national contexts is worthy of study. One aim of this monograph series is to facilitate cross fertilization of ideas and experience.

It is hoped that each issue will stimulate thought and discussion among informed persons who are actively involved in educational development and evaluation. By exchanging ideas and experiences, workers in different educational systems can do a great deal to increase their effectiveness.

Intending contributors are advised, in the first instance, to submit an outline of their proposed monograph to either of the co-editors.

CONTENTS OF PREVIOUS ISSUES: EVALUATION IN EDUCATION

CONTENTS OF CO-ISSUE: STUDIES IN EDUCATIONAL EVALUATION Vol. 8, No. 2