The Seventh International Conference
on Technology and Education

Brussels, Belgium March 20-22, 1990

A FRESH LOOK ON CONFIDENCE MARKING

D. LECLERCQ and Ph. de BROGNIEZ
University of Liège. Belgium

A. THE RATIONALE.

1) The social status of ignorance and doubt

We ignore a lot of things. Social life is organized according to a large «spreading» of knowledges : lawyers, surgeons, chemists, physicians know things others do not ... Common sense recommands us not to act by ourselves in a domain where we are incompetent.

We become dangerous, for ourselves and our neighbours, when we do not confess our incompetency (in piloting an airplane or knowing a foreign language, the effects of a drug, the starting time of a train, etc ...). If someone asks us for a piece of information we do not have, it is easy and efficient to confess directly our ignorance. The other person will ask to someone else. Know what we do not know is a precious thing.

Bertrand RUSSEL noted that «the problem in this world is that idiots are sure of themselves and wise persons full of doubts».

2) The learning issue.

Learning to learn consists first in being aware of what we know and what we do not know, then in managing to use the most appropriate means to get this lacking knowledge. In this respect, DESCARTES (1628) considered doubt as the motor of knowledge.

For all those reasons, self assessment is one of the most important objectives of the curriculum. The procedure we suggest is a general tool to contribute to this educational goal.

3) Theoretical basis.

Fundamentally our approach is based on CHOPPIN'S (1971) model 3 of mental activity happening in the mind of someone who answers a multiple choice question :

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

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.»

4) Variety of suggested methods.

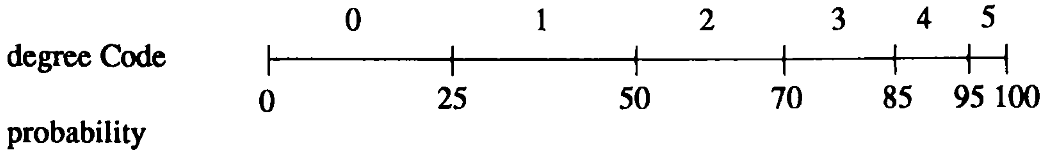
- DE FINETTI (1965) suggested a graphic mode of answering (nowadays promoted by BRUNO and BAXTER, 1989).
- SHUFFORD (1966) suggested a 10 points probabilistic scale. The ten confidence zones are equal and tariffs are computed according to decision theory (VON NEUMANN & MORGENSTERN, 1947; VAN NAERSSEN, 1966; LORD, 1970).
- Several other modalities have been tested but lacked theoretical basis related to the instructions and computation of appropriate tariffs (i.e. payoffs or reinforcements or scoring formulas).
- Our instructions are based on CHOPPIN's, DE FINETTI's and SHUFFORD's theoretical contributions . Our tariffs are computed according to decision theory so that he student is interested not only to estimate realistically his/her degree of doubt but also to report it without any biais.

B. CRUCIAL METHODOLOGICAL CONDITIONS

1) The instructions must offer a metric scale

Most researchs have used vague (ordinal) instructions such as «Tell me whether you are STRONGLY sure, FAIRLY sure, WEAKLY sure « about your answer. Ordinal processing of data are spurious since we have no garantee that even into a single person those «yardsticks» keep the same meaning from test to test, from item to item. Comparisons between different persons are, of course, to be excluded for the same reasons.

Here are instructions specifying codes to designate defined portions of the probability scale. Experiments show that (like in differential tresholds in psychophysics), sensivity is not the same at the extreme portion of the scale than in the middle.



Tariffs :	0	1	2	3	4	5
correct	13	16	17	18	19	20
incorrect	4	3	2	0	-10	-20

2) Tariffs must be computed according decision theory.

It is not sufficient to score with higher points a correct answer with a high confidence degree than a correct answer with a low confidence degree. All points must be computed so that the learner is interested in expressing his doubt with realism, without biais. The series of tariffs is an example of such a scale, that insures local optimality of each confidence degree.

3) We must distinguish measurements from payments

Numerous researches, most of which were published in the famous Journal of Educational Measurement, raise the question : «Are new (total) test scores (computed with new scales of tariffs) more valid and more reliable than classical ones (number of correct answers) ?»

Results from these experimental researches are confusing. Half of them show an increase in validity and a decrease in reliability whereas the others find the contrary ... without being able to explain why.

Actually, the problem itself is wrongly stated since the new total score is not a measure, but the combination of two different measures :

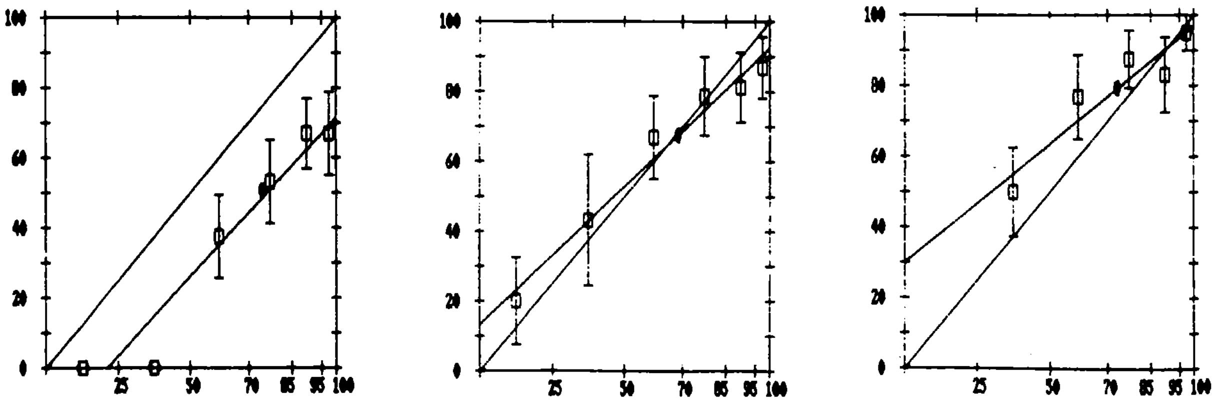
- the measure of ability (number of correct answers).
- the measure of realism (quality of self-assessment).

The new score can be more valid (i.e. reflect more accurately the learner's competency) only if he/she is realistic !

C. THE "CERT" SOFTWARE.

The CERT computer program computes several mathematical indices and outputs a graphic representation of realism. The following graphics present 3 examples of students' «realism» when using instructions hereover.

The student on the left overestimates himself, the student of the right underestimates and the student of the middle is very well calibrated.



Each little « square» represents a rate of success for the given portion of the axis, i.e. for the given confidence zone. Each square is accompanied by the interval of confidence (standard error of measurement) of the percentage it represents (formula $=\sqrt{pq/N}$)..

If the error range (drawn over and under the square) overlaps the diagonal line (that represents perfect realism), we cannot conclude that the student is un realistic.

D. THEORETICAL ISSUES

We have shown (LECLERCQ, 1983) that only the measurement of partial knowledge and of doubt makes possible the assessment of the informative value of an event, of information processing and judgement revision behaviors (bayesian approach).

It is the only fair way to score students (provided scores are adapted for «severity») and to inject a new source of knowledge about the individual learners in the training process.

E. REFERENCES

BRUNO J. & BAXTER J. (1989), An Application of Information Reference Testing, in Proceedings of the Sixth International Conference on Technology and Education, Orlando, vol. 2, 191-192.

CHOPPIN, B., (1970), An IEA Study of Guessing. A Proposal, Stockholm, International Association for the Evaluation of Educational Achievement. Unpublished Memorandum, IEA/TR/9.

DE FINETTI, B. (1965), Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item, Brit. Journ. of Mathem. and Statst. Psych., 18, 87-123.

DESCARTES, R. (1628), Règles pour la conduite de l'esprit.

LECLERCQ, D. (1983), Confidence marking, its use in testing, in B. CHOPPIN and N. POSTLETHWAITE, Evaluation in Education, An International Review Series, vol. 6, number 2, 161-287, Oxford : Pergamon.

SHUFFORD, ALBERT et MASSENGILL, N.E. (1966), Admissible Probability Measurement Procedures, Psychometrika, 31, 125-145.

VON NEUMANN, J. et MORGENSTERN, O., (1947), Theory of Games and Economic Behavior, Princeton Univ. Press.