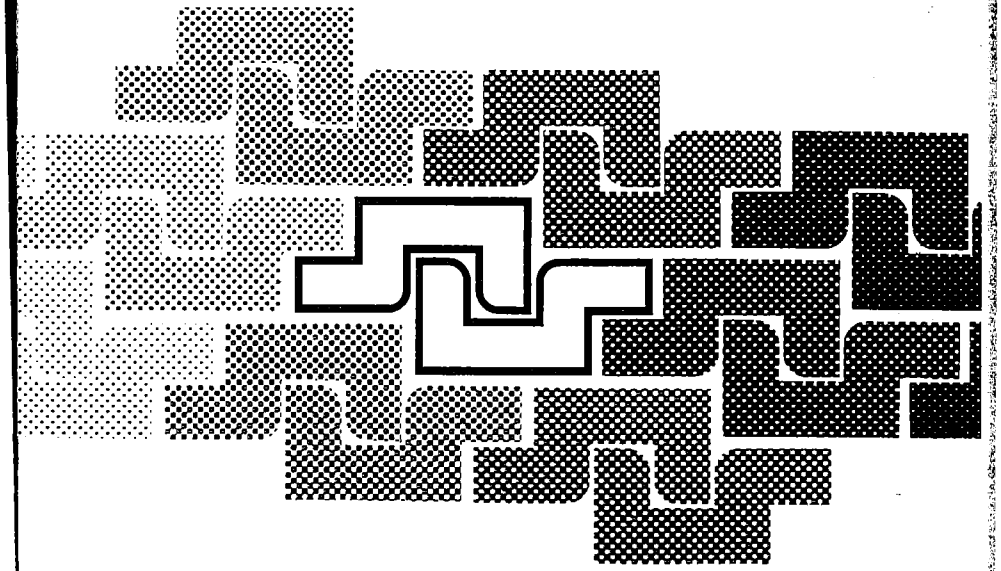


NATO ASI Series

Springer-Verlag
Berlin Heidelberg New York London Paris Tokyo
Hong Kong Barcelona Budapest

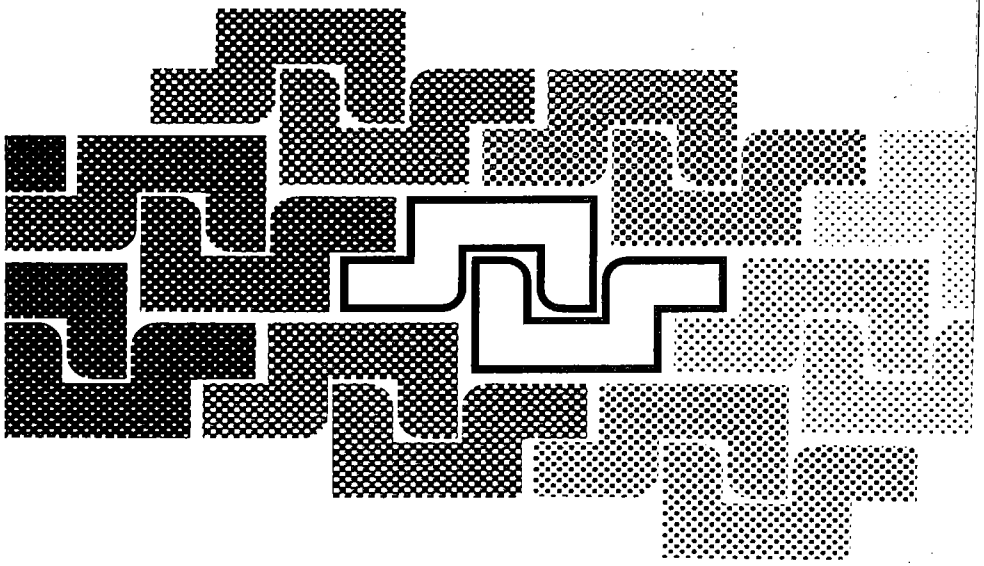
ISBN 3-540-56653-8 · ISBN 0-387-56653-8

ASI

D. A. Leclercq
J. E. Bruno

Item Banking: Interactive Testing and Self-Assessment

F 112



Item Banking: Interactive Testing and Self-Assessment

Edited by
Dieudonné A. Leclercq James E. Bruno

NATO ASI Series

Series F: Computer and Systems Sciences, Vol. 112

Validity, Reliability, and Acuity of Self-Assessment in Educational Testing

Dieudonné Leclercq

Université de Liège, Service de Technologie de l'Education, Batiment 32, Sart Tilman, 4000 Liège 1, Belgium, Tel. (32 41) 56 20 72, Fax (32 41) 56 29 44, e-mail: U017801 at BLIULG11

Abstract: When teachers use confidence marking, they should be aware that confidence estimation and confidence expression are influenced by a series of factors. Some of them have been studied in detail, such as the general human capacity to estimate one's knowledge (how far can people be sensitive, reliable and valid in appreciating their uncertainty).

This paper indicates how some of these factors have been studied, the results and the implications for designing test instructions, proper scoring rules and indices of the quality of self assessment.

Keywords: Subjective probabilities, self assessment, validity, reliability, sensitivity, confidence marking

The expression of a student's degree of confidence depends upon two series of factors: factors affecting the confidence *estimation* and factors affecting the confidence *expression.*

A. Factors affecting the confidence estimation

1. The student's ability

The student's ability should be validly and reliably reflected by the confidence degree.

The degree of ability could influence the student's judgment in two respects:

a) Students with perfect (100% correct) ability are likely to underestimate (it is impossible to overestimate). The reverse is true for student with 0 ability level.

b) Sensitivity is likely to be more subtle in the extremes (low probabilities and high probabilities) of the scale, as Edwards (1971) advocated by noticing that human estimators use ratio scales. Therefore, they can distinguish 95% (1 chance out of 20 to be wrong) from 99 % (1 chance out of 100 to be wrong) but not 50 % (1 chance out of 2) from 54 % (1 chance out of 2,17). This fits with what we know about psychophysics: sensitivity (acuity) is not in a linear, but in a logarithmic progression with the probability scale.

2. The human average capacity of self-estimating

a) The human sensitivity (or acuity or granularity).
We know we are unable to make a difference between two confidence degrees as close as 37% and 38%. But how far can an average person distinguish on a probability scale? How many different degrees? Is Miller's (1956) "magical number seven plus or minus two" applicable in this domain?
b) The human realism (the validity problem): how far is it reasonable to expect an average person to self-estimate? Does it vary from content to content? How many tests do we need to assess the validity of self-estimation ? Does it depend upon the number of degrees?
c) The human stability in time (the reliability problem): does it depend upon the numbers of degrees?

Obviously, those three concerns are interrelated and we shall provide data that answer several questions in the same time.

3. The personal sensitivity, reliability and validity

In this domain as in others in psychology, it is likely that interindividual differences occur, at least with untrained persons.

4. The content on which estimations are made

The content on which estimations are made could produce intrapersonal variations (a person may be more realistic in some domains than in others). As an answer to one of our questionnaires, testees noted that "they would express their probability differently if they were risking their life". There also exist domains where a bit of irrealism helps overcome difficulties of life, such as the 40-year old man saying: "I have already lived one third of my lifetime!"

5. The degree of familiarity and of training with the specific procedures used

As shown in Section J, self estimation can be inproved by practice, especially if the learners are exposed to the consequences (payoffs) of their actions, in an operant conditioning way.

B. Factors affecting the confidence expression

There can be a difference between what a person believes and what he/she expresses (says or writes) for a series of reasons.

6. The qualitative aspect of instructions

The qualitative aspect includes such effects as that of using different types of scales, for example,

- An ordinal scale (weakly sure, fairly sure, strongly sure), i.e. the more widely used instructions (unfortunately!).

- A continuous confidence marking system (De Finetti, 1965a) where the student expresses his/her confidence with any precision he/she likes (e.g. 0.3 as well as 0.318), with the computer applying a scoring function (see details in Leclercq, 1983).
- A 10-stars system (Michael, 1968) where the student has to distribute the ten stars (each representing 10 % confidence) among the alternatives. From his experimental data, Michael concluded that a classicaly scored test should be 1.7 times longer to reach the reliability obtained with a system of ten stars.
- A 5-stars system (De Finetti, 1965a)
- Locating the answer within a (triangular) space (De Finetti, 1965; Bruno, 1993)
- A choice between some precoded zones of the probability scale (Leclercq, 1983 and Leclercq et al., 1993).

Although they are profusely used, ordinal scales lead to almost uninterpretable data. We all should consider "admissible probability measurement procedures" (Shuford et al., 1966) for assessing partial knowledge.

7. The consequences often represented by the scale of tariffs

Inappropriate scales could elicit lies to optimize the total amount of points, i.e. the final score. The scales of tariffs should, consequently, have been constructed carefully, according to decision theory (Van Naerssen et al., 1965) and be kept clear for the student so that he can compute its own score himself (Raiffa, 1970; Leclercq, 1983, p. 207).

8. Personal attitude towards risks

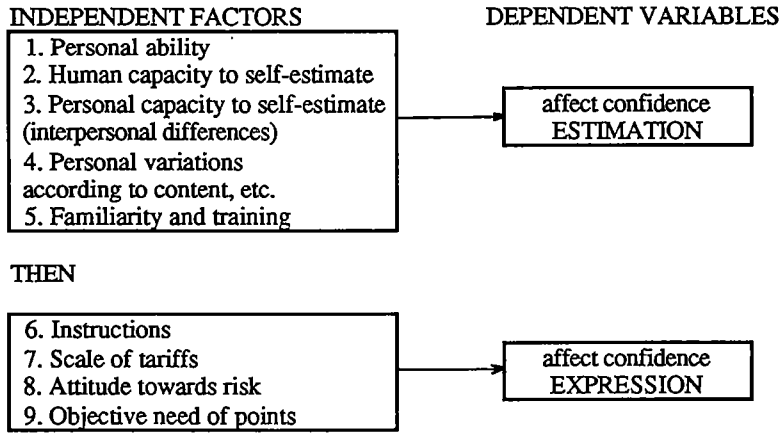
Atkinson (1971) has described persons motivated by the "hope of success" and others by the "fear of failure". Coombs has described (in his "unfolding theory") preferences for amounts of risks and of probabilities. The problem of conservation of expected utility can be formulated in the classical piagetian way: some are non conservants; for instance they are overinfluenced by the consequences, others by the probabilities (they do not apply the compensation principle).

9. Need for points

Students having few points and wanting to maximize their gains in order to reach the pass/fail threshold could be inclined to adopt a more cautious behaviour than people that know that their total score will anyway be beyond the passing score. Van Naerssen et al. (1966) have studied this problem, concluding that there is no reason for rejecting the assumption of linearity in the students' motivation of maximizing their total score.

C. The general picture of factors affecting confidence degrees

Those considerations lead to the following schema:



Factors 3, 4 and 5 are sometimes refered as personality factors and there is a reluctancy in "melting" them with the assessment of knowledge. As will be shown in sections I and K hereafter, individual realism can be "computed" and its weight in the score fixed at will.

Teachers are interested in eliciting the best student's estimate and having it expressed without bias.

The following lines will provide data on several of the components of the first part of the schema hereover. Other publications (Leclercq, 1983, 1990, 1993) have been dedicated to the second series (6, 7, 8, 9) of factors.

D. An experimental approach

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

1. The Sentence Confidence Guessing Game (SCGGame)

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 guess other letters until he finds the correct one. The experimental data are presented by indicating below each letter the number of trials needed until the correct answer was found.

We slightly changed this game by allowing the student to provide only one guess (a letter) and by asking him/her to accompany his/her answer with a confidence degree. The player is provided with the correct answer after each trial. We will refer to this as the Letter Confidence Guessing Game (LCGG).

For the validity/reliability/sensitivity experiment, we used a "Sentence" Confidence Guessing Game (SCGG), in which a long text (about 3 pages) is chosen from a book. The odd lines of the text are printed and truncated whereas the even ones are not. The subjects are asked to predict the first letter of the truncation.

Cutoff points have been chosen in order to obtain items of various difficulties (ideally, a rectangular distribution with an average mean of 0.50). Figure 1 shows the distribution of the facility indexes for the 100 questions in the experiment.

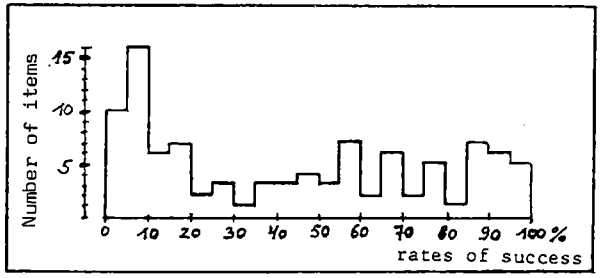


Figure 1

2. The three confidence degrees scales

Figure 2 presents an example of such an item. Here, the correct answer is L, since the truncated text is "The magical number seven plus or minus two." 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.

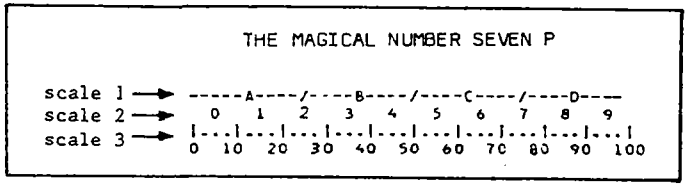


Figure 2

3. The tariffs

The subjects were told that "points would be given in such a way that, if they want 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 Figure 3). The maximal score is +50 (TC for confidence degree 100 on scale 3 of Figure 2) whereas the minimal score is –100 (TI for confidence degree 100 on scale 3).

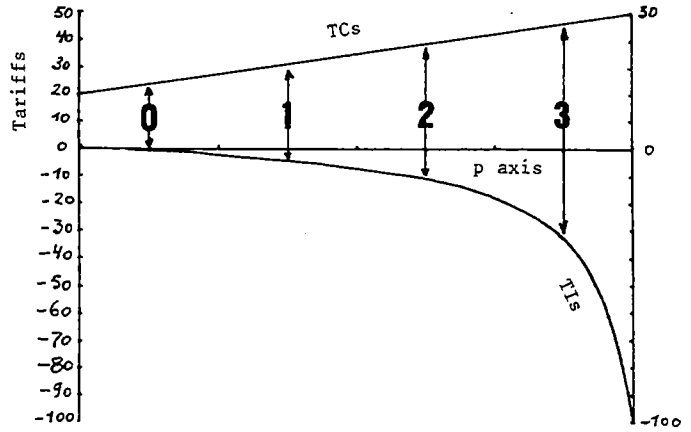


Figure 3

4. 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 received the correct answer 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, or his/her 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: 4 possible confidence degrees (25% each);
   - Scale B: 10 possible confidence degrees (10% each);
   - Scale C: 40 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. Furthermore, on retest, subjects were invited to describe the way in which they chose their degrees of confidence.

5. The general results

The whole test-retest procedure was completed by 121 subjects. Figure 4 shows that 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.

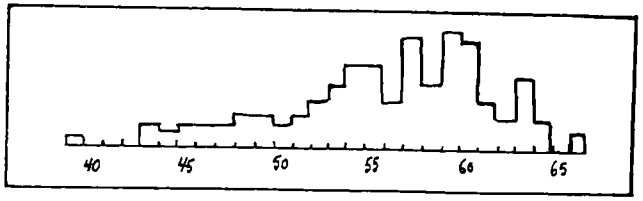


Figure 4

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. Therefore, the subjects could have estimated their confidence degrees less carefully, dedicating less time to this process.

E. Results concerning stability

1. The basic data

For each subject, a computer program prints:
- The scatter plot of the 100 confidence degrees used on the test (t1) and on the retest (t2);
- The two means (M1 and M2) and standard deviations (SD1 and SD2) of those degrees;
- The Bravais Pearson correlation coefficient (R) between the t1 and the t2 degrees;
- 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 Figure 5.

2. The results

Here are two examples (for subjects 104 and 94) of the two histograms (before and after test and retest) with a rather good replication pattern (a correlation of 0.80).


```
YOUR CODE = 104
(BEFORE) 40 DEGREES
 3 *
 7 *
11 ***
15 ***
19 **************
23 ****
27 ************
31 ********************
35 ******************
39 **************************
YOUR CODE = 104
(AFTER) 40 DEGREES
 3 *
 7 *
11 *
15 ***
19 ****
23 ************
27 ***************
31 ********************
35 ***************
39 *************************
```

```
YOUR CODE = 94
(BEFORE) 40 DEGREES
 3 ******************************
 7 *************
11 ***************
15 ****
19 **************
23 ******
27 **
31 *********
35 ***
39 ******
YOUR CODE = 94
(AFTER) 40 DEGREES
 3 ******************************
 7 *************
11 ****************
15 *******
19 ***********
23 ****
27 ******
31 **********
35 ****
39 *
```

Figure 5

For the 121 subjects, the Median correlation observed is 0.56 (not so bad for a first trial by untrained subjects).

```
                                        x
                                  x     x          x
                                  xx    x    x  x x    x  x
                                  xx    xxx xx xxx x    x  x
                                  xx    xxxx xxxxxx x x xxx x
                  x  x x xxxxxxx  xxxxxxxxxxxxxxx x  xxxxx
x   xx    x   xxxxxxxxxxxxxxx     xxxxxxxxxxxxxxxxxxxxxxxxxxx   xx

. 20      . 30      . 40      . 50      . 60      . 70      . 80
                                  Median
```

F. Results concerning acuity

1. Spontaneous uses of the 40 degrees

Whereas 40 degrees could be used, only 11 out of them have been systematically used, may be because the preprinted formula sheet was not attractive for "intermediate degrees".

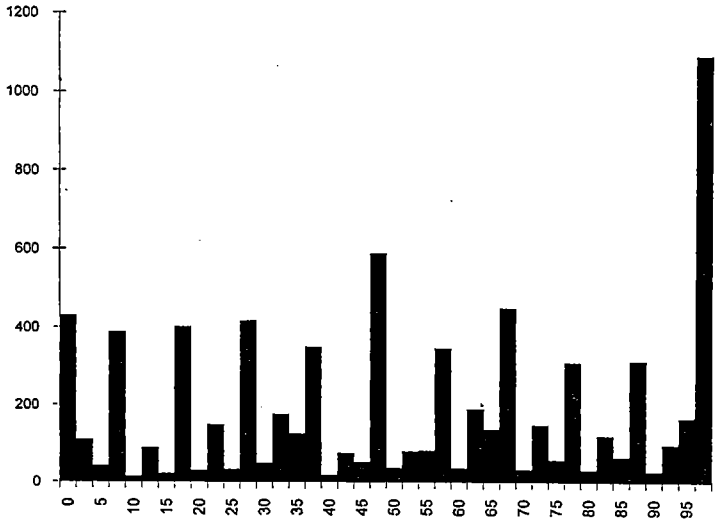


Figure 6

2. The Replication Histograms (or profiles)

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

The replication histogram of a given degree (say X) is established according to the following principles (see Figure 6).

- 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 on the retest at the very place where degree X was used in the test. Actually, when degree X has been used in a test, degrees close to degree X are used in the other test (and degree X should be the most used of all of them).

In order to make the graphs clear, the tops of the histogram rectangles have been joined. Only these "profiles" are presented. The replication profiles computed from the 40 most realistic subjects for the 4 degrees of the rawest scale (A) are presented in Figure 7 (confidence degrees 0 and 3) and Figure 8 (confidence degrees 1 and 2).

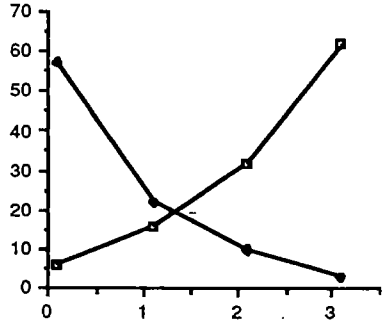


Figure 7

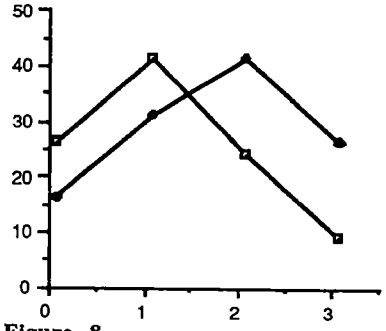


Figure 8

As can be seen, the top of the replication profile for degree 0 is 0, it is 1 for degree 1, etc. Our adult untrained subjects seem to have had no problem in dealing with 4 degrees (scale A) and it is likely that they can handle more sophisticated scales.

Figure 9 presents the 10 replication profiles for the B scale for 34 persons. It appears that the mode for X is X except for confidence degrees 7 (mode is 8), and 3 (mode is 4), and that for degrees 2, 4, 5 and 6 the peak is not neat!

This "overlap of some degrees" seems to indicate that 10 degrees is too much, either for untrained individuals or for this kind of work.

It seems also obvious that replication is better at the extreemes (low side or top side of the scale) than in the middle of the probability scale. Edwards (1967) has provided a possible explanation about that: "Human beings think not in probabilities but in ratios or odds. For instance, 99% is 1 chance out of 100 of being incorrect, whereas 95% is 1 chance out of 20 and 90 % is 1/10. Those ratios are quite different from each other. Whereas in the center of the scale 50% (1/2) is too close to 46 % (1/2.17).

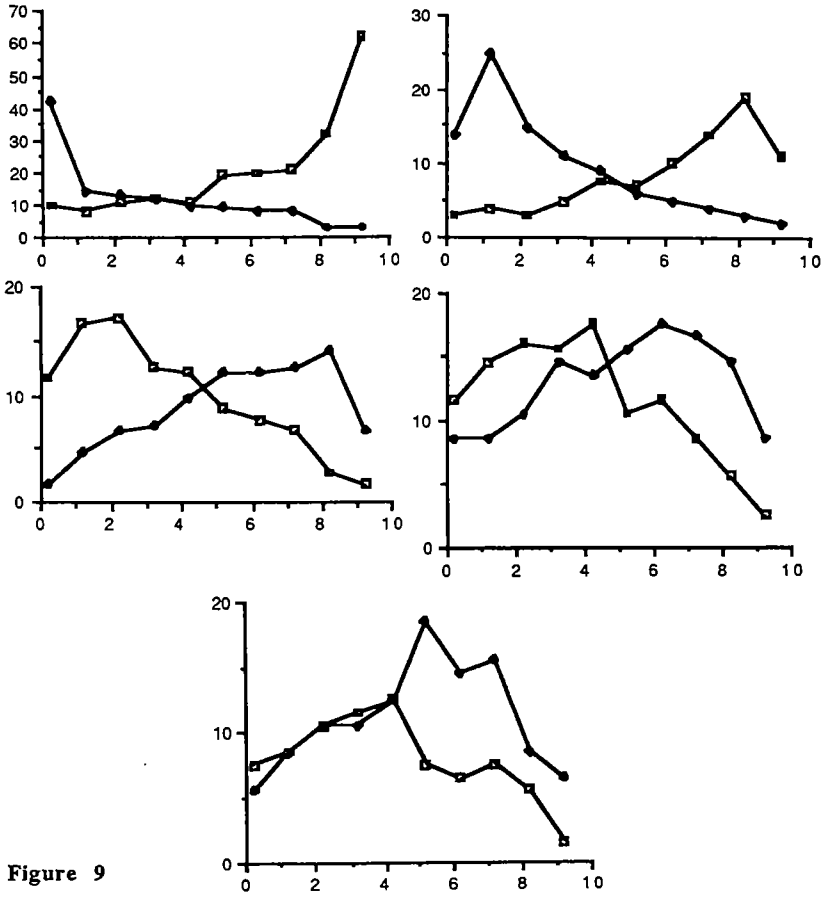


Figure 9

3. Subjects' introspection

Various questions have been asked to the subjects in order to know how they chose a given degree of confidence.

a) *Which sequence do you follow when you use the scales to indicate your confidence degree?* (4, 10, 40) or (40, 10, 4) or (10, 4, 40), etc. ?
Here are the answers in decreasing order of importance (number of observed cases):

1.	4, 10, 40	(56)
2.	40, 10, 4	(18)
3.	10, 40, 4	(15)
4.	10, 4, 40	(11)
5.	40, 4, 10	(10)
6.	4, 10 or 40	(9)
7.	40, 10 or 4	(7)
8.	10, 40 or 4	(6)
9	4, 40, 10	(3)

Starting with the 4 degrees scale (top down method) is used by 68 persons (50%).
Starting with the 40 degrees scale (bottom up method) is used by 35 persons (26%).
Starting with the 10 degrees scale (intermediate) is used by 32 persons (24%).

b) *How do you come up with a specific degree?*

Here is a selected series of answers:
- When I am perfectly confident, I start from 100% and when I am not confident at all, I start from 0%.
- I point a given place on the line (from 0 to 100) with no concern for any numerical value.
- For me there are three situations : sure, doubt (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 my confidence degrees for the answers I was perfectly confident (100%), then to the rest of the items.

c) *On what does the choice of a confidence degree depend?*

The 135 interviewed persons answer as follows:
28%: mainly on my subjective probability AND subsequently on risk.
26%: only on my subjective probability.
19%: on my subjective probability and on risk, equally.
18%: on risk only.
9%: mainly on risk, and subsequently on my subjective probability.

The popularity of the three first propositions (73 %) is a good indicator for the validity of the whole procedure. Some subjects noted that their strategy depends upon the situation (circumstances and objectives): if they were to risk their life, their strategy might change.

d) *What is your ideal number of degrees on the probability scale?*

10 degrees:	41
20 degrees:	13
40 degrees:	8
100 degrees:	4
10 to 20 degrees:	3
4 degrees:	3
50 degrees:	2

Others (5d., 6d., 7d., 8 d., 7 to 10 d., 4 to 10 d., etc.)

Some subjects that have chosen 10 degrees as the ideal scale added interesting comments:
-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%.
A subject explained that he has chosen the 20 degrees scale because he used it in scholar setting.

Other interesting comments were made:
- 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 39%)?
- This game can be learned and subject could improve their ability.
- I would prefer to give several answers and give to each a probability (this teacher was a linguist).

In general, subjects pointed out the importance of the number and the kind of items, of the situation (real consequence or not) of the subject (if he is "words minded" or not) and of the content of the text.

4. Conclusions for acuity

Conceiving a perfect test-retest situation is a hard job. To insure a same level of uncertainty on the test and on the retest, subjects must not reformulate the hypothesis about the correct answer and, consequently, change the probabilities.

We have not succeeded in building a situation where subjective probabilities are expressed given fixed hypothesis. Such a conditional game should be developed.

From our Confidence Guessing Game (CGG), it has been possible to observe reasonable test-retest stability.

On the resolution side, objective data as well as subjects' opinion show that for this kind of game, for this kind of (untrained) graduate adults, the maximal number of degrees was between 6 and 8. This result corroborates G.A. Miller's opinion that our spontaneous resolution in perceptual domains is "the magical number seven, plus or minus two".

Moreover, we can formulate the hypothesis that our "sensibility" (or resolution or acuity) is better in some portions of the probability axes (the extremes) justifying W. Edwards' procedure (1967) of using odds (that have logarithmic properties).

All those observations are of interest for conceiving an optimal scale to be used in school settings (Leclercq et al., 1992). There must only be a few degrees for practical reasons, but how many exactly and where on the axis of probabilities?

The present study can inspire new experiments to answer more precisely those questions.

G. Results concerning validity

1. With the 10 degrees scale

The calibration curve of 20 selected persons highly realistic (see Figure 10) shows the ambiguity between degrees 3 and 4 and between degrees 5 and 6.

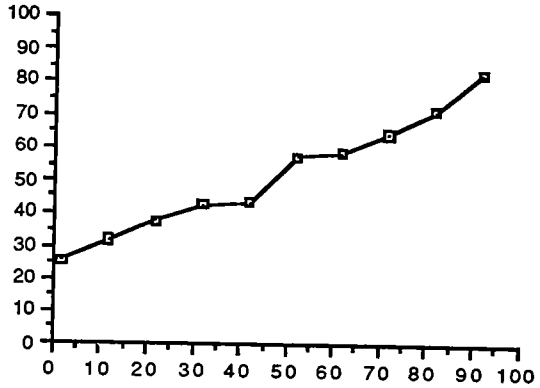


Figure 10

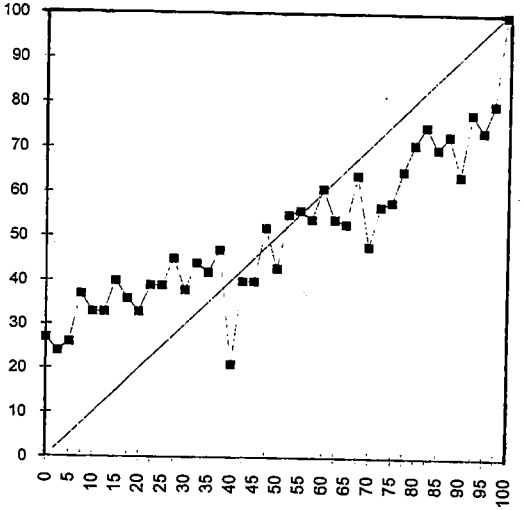


Figure 11

Things happen as if people had their 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 subjects had difficulties in handling 10 degrees.

2. With 40 degrees scale

It becomes impossible to build replication profiles, because some degrees have not enough observations. Figure 11 presents the general calibration-curve, where reliable values (computed on sufficient data) are represented by dark dots.

This calibration curve reinforces the impression (already given by the "acuity results") that adults cannot, obviously self-estimate validly with 10 degrees, but probably with three or four less, coming close to Miller's famous "magical number seven".

3. Ceiling and floor effects

A highly competent person (who succeeds in almost 100% of the occasions) can hardly overestimate! The same for a very incompetent one (succeeding close to 0% of the trials): he/she can hardly underestimate. Therefore, (normal or small) underestimation is to be expected for the former and overestimation for the latter. This explains that the slope of the calibration curve is lower than 1 (the diagonal line).

H. Formulas to compute the indices

We have developed the following formulas, based on previous research (Brier (1950), Adams & Adams (1961), Oskamp (1962), Murphy (1972, 73, 74) and Lichtenstein et al. (1977)), we have described elsewhere (Leclercq, 1983):

Acuity (or subtlety): The tendency to diversify one's judgments.
Formula: Standard deviation of the various rates of successes of the various degrees used.

Centration: Difference between the rate of (objective) success and the average confidence degree (result is either positive, negative or zero - perfection).
Formula: Difference between Average central values of confidence degrees used (or average confidence) and Rate of correct answers (percentage).

Coherence: The tendency of the successive rates of correct answers to fit a straight line (even if it is not the diagonal).
Formula: Correlation (Bravais-Pearson) between rates of successes and central values of the successive zones (when the degree has been used).

Realism: The tendency of the rates of correct answers to fit the diagonal.

Formula:

$$ERR = \sqrt{\frac{\sum_i (RS_i - CV_i)^2 \cdot NU_i}{NA}}$$

where RSi = Rate of Success of confidence degree i.
CVi = Central Value of confidence degree i.
NUi = Number of uses of confidence degree i.
NA = Number of answers.
ERR = a quadratic expression of error of realism.

REALISM = 1 – (5 ERR)

I. The need for norms

ACUITY is ...	if it is	% of 311 students
IDEAL	= 50 %	0 %
EXCELLENT	superior to 25 %	14 %
GOOD	between 23 % and 24 %	12 %
SATISFACTORY	between 20 % and 22 %	30 %
WEAK	between 16 % and 19 %	27 %
INSUFFICIANT	less than 16 %	15 %
MINIMAL	= 0 %	2 %

CENTRATION is ...	if the difference is	% of 311 students
IDEAL	= 0 %	1 %
EXCELLENT	less than 3 %	18 %
GOOD	from 3 to 6, 99 %	19 %
SATISFACTORY	from 7 to 10, 99 %	19 %
WEAK	from 11 to 15, 99 %	17 %
INSUFFICIANT	16 % and more	26 %
MINIMAL	97.5 %	0 %

COHERENCE is ...	if the correlation is	% of 311 students
IDEAL	= 1	0 %
EXCELLENT	superior to 0.97	13 %
GOOD	between 0.93 and 0.96	18 %
SATISFACTORY	between 0.85 and 0.92	28 %
WEAK	between 0.75 and 0.84	18 %
INSUFFICIANT	less than 0.75	23 %
MINIMAL	- 1	0 %

REALISM is ...	if the index is	% of 311 students
IDEAL	= 1	1 %
EXCELLENT	superior to 0.95	15 %
GOOD	between 0.91 and 0.94	18 %
SATISFACTORY	between 0.84 and 0.90	22 %
WEAK	between 0.70 and 0.83	22 %
INSUFFICIANT	less than 0.70	22 %
MINIMAL	- 3.75	

Those values are just provisional references. Concurrent and predictive validity measures (with achievement and grades) are currently computed in experimental settings. Simple correlations could be misleading since overestimation (in uncompetent persons) may be worse than underestimation (in competent persons) whereas their realism index may have the same values.
In addition, intrapersonal variations of realism due to different contents needs also to be explored, when familiarity (and subsequent quality of realism) with the procedure has been reached.

J. Can realism be improved by practice?

With the help of F. Lambert (1992), we have studied the evolution of these 4 indices for (86) students involved in four settings:

1. A home answered test (March 92). It is supposed to be easy since no constraint of time exist and the students could discuss and cooperate with each other to answer the test.
2. A first computer test (end of April 1992). It is supposed to be the most difficult since each student has to answer alone, with time pressure.
3. A second computer test (May 1992). Easier than the previous one since familiarity with the computer testing situation has developed.
4. The last test (beginning of June 1992), paper administered. The easiest since familiarisation to the whole procedure has occurred, there is no time pressure and the students have an overview of all the questions (since it is on paper).

Here is the evolution of the average values of the 4 indices:

	Acuity	Centration	Coherence	Realism
Max	35	0	1	1
1.Home(March)	18.8	13.4	0.38	0.56
2.Comp.(April)	22.4	17.5	0.31	0.40
3.Comp.(May)	21.5	16.5	0.45	0.52
4.Paper (June)	21.5	10.6	0.78	0.80

Centration, Coherence and Realism have improved continuously during the three "real" tests (i.e. really valued for the final score), i.e. tests 2, 3 and 4.

K. Conclusions

Conceptual and technical tools exist to assess students' acuity, centration, coherence and realism of self-estimation of competency. Norms are developing. This behaviour can be trained. It is worth being explored further.

Just like Bruno De Finetti (1965b), we believe that

"Partial information exists. To detect it is necessary and feasible" (p. 109)

and that

"It is only subjective probability that can give an objective meaning to every response and scoring method" (p. 111).

This meaning has not always been well interpreted. For instance, scoring has been confused with measurement. The previous issues resulted in a first wave research proved to be a dead end in the late seventies. Nowadays, a new wave is developing rapidly. Prototypes of this movement are Shuford's, Hunt's and Bruno's work as well as the TASTE Approach, pieces of which can be found in Leclercq & Bruno (1993).

References

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

Attneave, F. (1959). Application of information theory to psychology. New York: Holt, Rinehart and Winston.

Brown, T.A. & Shuford, E.H. (1973). Quantifying uncertainty into numerical probabilities for the reporting of intelligence (Report R-1185-ARPA), Santa Monica, Cal.: Rand Corporation.

Bruno, J. (1993), Using testing to provide feedback to support instruction: a reexamination of the role of assessment in educational organizations. In: D. Leclercq, J. Bruno (eds.): Item banking: interactive testing and self-assessment. NATO ASI Series F, Vol. 112. Berlin: Springer-Verlag (this volume).

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

Coombs, C.H., Dawes, R.M., Tversky, A. (1970). Mathematical psychology. Englewood Cliffs NJ: Prentice Hall.

De Finetti, B., (1965a). La décision et les probabilités, Revue des Mathématiques pures et appliquées, Bucarest, 405-413.

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

Edwards, A.L. (1967). Statistical methods. New York: Hold, Rinehart and Winston, 2nd edition.

Edwards, W., (1967), Probabilistic information processing by men and man-machine systems, in La simulation du comportement humain, Paris, Dunod, p. 187.

Hunt, Darwin P. (1993), Human self assessment - theory and application to learning and testing. In: D. Leclercq, J. Bruno (eds.): Item banking: interactive testing and self-assessment. NATO ASI Series F, Vol. 112. Berlin: Springer-Verlag (this volume).

Leclercq, D. (1975). L'évaluation subjective de la probabilité d'exactitude des réponses en situation pédagogique. Thèse de doctorat en Sciences de l'Education, Université de Liège, Institut de Psychologie et des Sciences de l'Education.

Leclercq, D. (1983), Confidence marking, its use in testing. In: Postlethwaite, Choppin (eds.) Evaluation in Education, Oxford : Pergamon, 1982, vol. 6, 2, 161-287.

Leclercq, D., Boxus, E., de Brogniez, P., Lambert, F., Wuidar H. (1993), The Taste approach: General implicit solutions in MCQs, open books exams and interactive testing. In: D. Leclercq, J. Bruno (eds.) Item banking, interactive testing and self-assessment. NATO ASI Series, Vol. 112. Berlin: Springer Verlag (this volume).

Leclercq, D. & de Brogniez Ph. (1990), A fresh look on confidence marking. In: Estes, Heene, Leclercq (eds.) New pathways to learning through educational technology. Proceedings of the Seventh International Conference on Technology and Education, Brussels, vol. 1, pp. 646-649.

Lichtenstein, S., Fischhoff, B., Phillips, L.D. (1975). Calibration of probabilities: the state of the art, decision making and change in human affairs. Proceedings of the Fifth Research Conference on Subjective Probability, Utility and Decision Making, Darmstadt, 1-4 September, D. Reidel.

Lindley, D.V. (1971). Making decisions. London: Wiley.

Luce, R.D., Raiffa, H. (1966). Games and decision. New York: Wiley.

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

Miller, G.A. (1956). The magical number seven, plus or minus two. Psychological Review 63, 81-97.

Murphy, A.H., & Winkler, R.L. (1974). Subjective probability forecasting experiments in meterorology: some preliminary results. Bulletin of the American Meteorological Society 55, 1206-1216.

Pitz, G.F. (1974), Subjective probability distributions for imperfectly known quantities. In: Gregg, L.W. (ed.) Knowledge and Cognition. New York: Wiley, pp. 29-41.

Raiffa, H. (1970). Decision analysis, introductory lectures on choice under uncertainty. New York: Addison-Wesley.

Savage, L.J. (1951). The foundations of statistics. New York: Wiley.

Shannon, C.E. (1951). Prediction and entropy of printed English. Bell Syst. Techn. J. 30, 50-64.

Shuford, E., Albert, A. & Massengill, N.E. (1966), Admissible probability measurement procedures. Psychometrika 31, 125-145.

Shuford, E. (1993), In pursuit of the fallacy: resurrecting the penalty. In: D. Leclercq, J. Bruno (eds.) Item banking: interactive testing and self-assessment. NATO ASI Series F, Vol. 112. Berlin: Springer-Verlag (this volume).

Van Naerssen R.F. & Van Beaumont, R. (1965). Ervaringen met een Zekerheidsaanduiding bij objektieve Tentamens. Nederlands Tijdschrift Psychologie 20, 208-315.

Van Naerssen, R.F., Sandbergen, S. & Bruynis, E. (1966), Is de Utiliteitscurve van Examenscores een Ogief? Nederlands Tijdschrift Psychologie 21(6), 358-363.