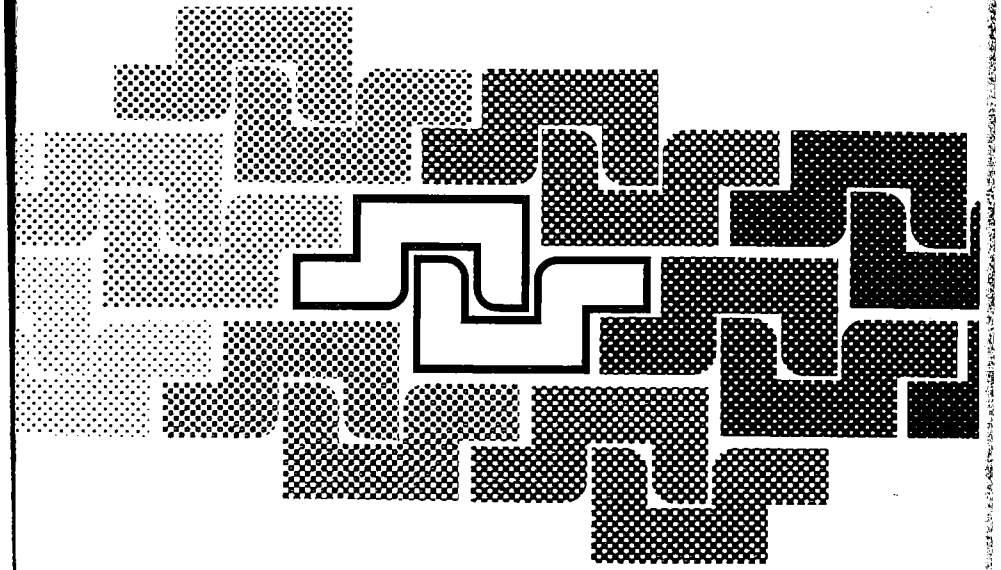


NATO ASI Series

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

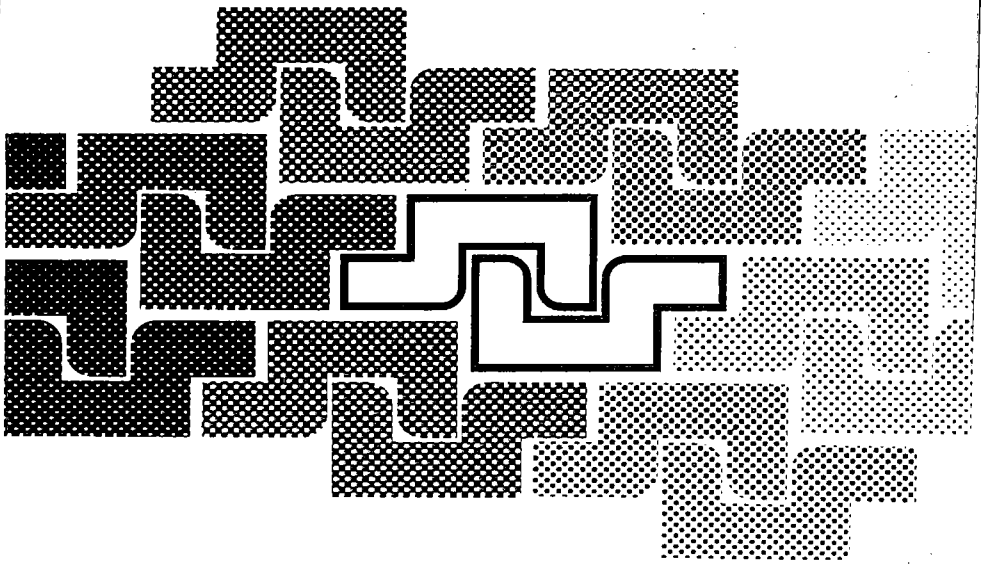
ISBN 3-540-56653-8 · ISBN 0-387-56653-8

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

The TASTE Approach: General Implicit Solutions in Multiple Choice Questions (MCQs), Open Books Exams and Interactive Testing

D. Leclercq, E. Boxus, P. de Brogniez, H. Wuidar, F. Lambert

Université de Liège, Service de Technologie de l'Education (STE), 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: The efficiency of school to efficiently train in higher order cognitive skills is sometimes questioned. Principles and techniques (such as open books exams, implicit questioning, self assessment) have been developed to enhance cognitive vigilance, data processing skills and metacognition. The details of those techniques are provided as well as some results about question characteristics and student opinions. Further developments are considered and perspectivized with ambitious training and assessment goals.

TASTE is an anonym that means "Towards an Adult System of Training and Evaluation". Its rationale and components will be described hereafter.

1 General Implicit Solutions in Multiple Choice Questions (MCQs)

1.1 The recall/recognize issue

It has been frequently advocated that MCQs do not assess the same type of ability as open questions or essays do. In particular, a student can *recognize* among the printed solution of a MCQ the correct answer whereas he/she would not be able to *recall* it.

It has been suggested (Wood, 1977) to add the option "none of those solutions" as the last solution. For the MCQs where this option is the correct one, the student cannot "recognize" but is forced to recall. We have adopted this principle, as well as the option "all of them", referring to Noizet and Fabre (1975) who present the possible solutions of a multiple choice question in a "simplex" structure.

1.2 A taxonomic challenge

It is well known that MCQs insidiously lead teachers to ask questions that deal only with *details*. The reason for this is that on details everybody agree and the correct answer will not be disputed.

Wood (1977) has suggested to use MCQs where the student could answer "I *lack data* to be able to answer" or "there is an *absurdity* in the stem that makes the whole question *nonsense*." These types of question force the student to consider the relevance of the data or of the question themselves, i.e. to *understand* the problem and *analyse* the way it is stated instead of limiting him/herself at plain *knowledge* and *application* levels (in the terms of Bloom's taxonomy of cognitive objectives). Whe have adopted this principle, making it systematic, i.e. applying it in all our questions.

1.3 The hidden curriculum

The hidden (or latent) curriculum can be defined as "what nobody teaches but everybody learns". It is true that, by law of effect (operant conditioning), feedbacks not only reinforce the selection of the correct answer but also the responding behavior.

In this way, through explicit questioning, students "learn" a series of hidden principles such as :

- You should answer when questioned (What about questions that do not deserve an answer or that should not be asked ?);
- You should wait to be questioned to answer (This lowers the tendency to raise questions spontaneously, a very healthy behavior in cognitive functioning).

Through "limpid" and simple questions, students learn that :

- When the authority (here the teacher) asks a question, it is always a relevant one and there always exists an answer, more specifically *one* answer and *only one* (Whereas we all are frequently presented wrongly formulated questions or questions for which there exists no available answer or others for which several answers can fit).

This hidden curriculum is totally in opposition with real life that requests from persons to "detect problems", to assess likelihood of reasoning, data necessity, data sufficiency, etc.

The gap is so great that only a *systematic and massive counter curriculum approach* can counterbalance the hidden curriculum's negative effects. *Implicit questioning* is an endeavour in this direction.

1.4 Implicit questioning

The so-called Dr. Fox's experiment is a game where a lecturer presents stupidities, nonsense reasoning, incorrect data ... up to the moment a person from the audience stops him/her, questioning the relevance of the content. Often, lecturers can speak for minutes and minutes before being interrupted: we have not been trained in *detecting* awkward reasoning, ill stated problems, etc.

Actually, the majority of the audience is *able* to decide it but *only if they are asked*. Therefore the training must focus of the detection side; that is the reason why we developed implicit questioning, i.e. *the teacher creates a situation* in which he hopes (he expects) the student will react *but without presenting it as a question.*

In our MCQ instructions, students are informed that they will have to consider two types of *solutions* :

- the *printed* ones;

- *additional ones* that have been described at the start of the test (and will remain under the student's eyes during the whole testing) but *that will NOT be repeated in each question* : the implicit solutions.

1.5 General solutions

We have limited the implicit solutions to four types that can be general, i.e. stand for each of the MCQs appearing in a school test. These types include:

1. REJECT (R or 6) : None of these proposed solutions is correct.
2. TOTAL (T or 7) : Each of these solutions is correct.
3. MISSING (M or 8) : The stem of the MCQ lacks a piece of information that is necessary to decide which solution is correct.
4. ABSURDITY (A or 9) : The stem of the MCQ contains an illogical statemant, i.e., an error, an absurdity that makes the whole question irrelevant.

As a general rule, the 9 answer (A) has priority over all the others : an absurdity has to be detected and pointed out *instead of (and before)* trying to answer the question. It is not sufficient to answer 6 (Reject : None) since, when the stem countains an absurdity, all solutions are automatically incorrect answers.

In the same way, the 8 solution (Lack of data) has priority over the 7 solution (Total) since, when data is lacking, all solutions are *potentially* correct.

Here are a few examples illustrating what has to be answered when those General Implicit Solutions are considered :

Question		Correct Answer
The capital of France is	1. Lille 2. Lyon 3. Paris	3
The capital of Italy is	1. Berlin 2. Praga 3. Tokyo	6
U.K. contains	1. England 2. Northern Ireland 3. Wales	7
How old was Lincoln ?	1. 20 years old 2. 30 years old 3. 40 years old	8
In which year did Napoleon and Hitler meet?	1. 1850 2. 1915 3. 1945	9

1.6 Advantages of General Implicit Solutions (GISs)

The Reject (6 or None) solution forces the student to answer mentally *before* having a look to the suggested solutions, *then* to check among them whether his/her own solution is presented. This transforms the usual "recognition" performance into a "recall" one.

The Total (7 or all) solution trains the student to consider that sometimes there exist several correct solutions to a given problem.

The Missing (or lack) of data (8) solution enables the teacher to assess Bloom's levels of cognitive processes such as analysis and intelligent application, since being able to apply a principle implies to be able to detect when it is impossible to apply it and why.

Last, the Absurdity (9) solution is a good way to assess "deep understanding", namely in detecting contradictions.

Usually, all those GISs are not introduced all together in the instructions for untrained students. A strategy of progressivity is recommended.

2 Confidence marking

It has been shown elsewhere (Leclercq, 1983) that faced to inappropriate scales of tariffs (point awards), students discover that some strategies consisting to bias their intimate estimation of confidence pay more than telling the truth.

In order to reinforce students to tell the truth, to make it the optimal behaviour, tariffs must be computed according to decision theory (Von Neumann & Morgenstern, 1947).

It must be noted that very few researchers have followed these methodological conditions, which has resulted in inconsistent data and finally ruined a fruitful approach, which was almost abandoned during the late 1970s. Fortunately, a few scholars have continued to promote strict methodological requirements, what De Finetti (1956) calls "Methods for discriminating levels of partial knowledge ...", what Shuford et al. (1966) call "Admissible probability measurement procedures", and what Van Naerssen (1962) calls "A scale for the measurement of subjective probabilities". To these pioneers must be added Darwin Hunt and James Bruno, the systems of whom are described in Leclercq and Bruno (1993).

Actually, five fundamental methodological conditions should be fulfilled to use confidence marking in a valid way.

2.1 The instructions must offer a metric scale

Most researches 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 guarantee that even within 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 specifiying codes to designate defined portions of the probability scale. Experiments show that (like in differential thresholds in psychophysics), sensitivity is not the same at the extreme portion of the scale than in the middle (Leclercq, 1983, 1992).

2.2 Tariffs must be computed according to 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 (positive and negative ones) must be computed so that the learner is interested in expressing his doubt with realism and without bias i.e. in telling the truth or admitting to their actual knowledge. The series of tariffs hereafter is an example of such a scale, that insures local optimality of each confidence degree (see Figure 1).

For the previous scales, the tariffs we adopted were as follows:

Confid. code:	0	1	2	3	4	5
	0 25	50	70	85	95	100
Tariffs:						
correct	13	16	17	18	19	20
incorrect	4	3	2	0	–6	–20

A series of other "acceptable" scales and tariffs have been described elsewhere (Leclercq, 1983, p. 210).

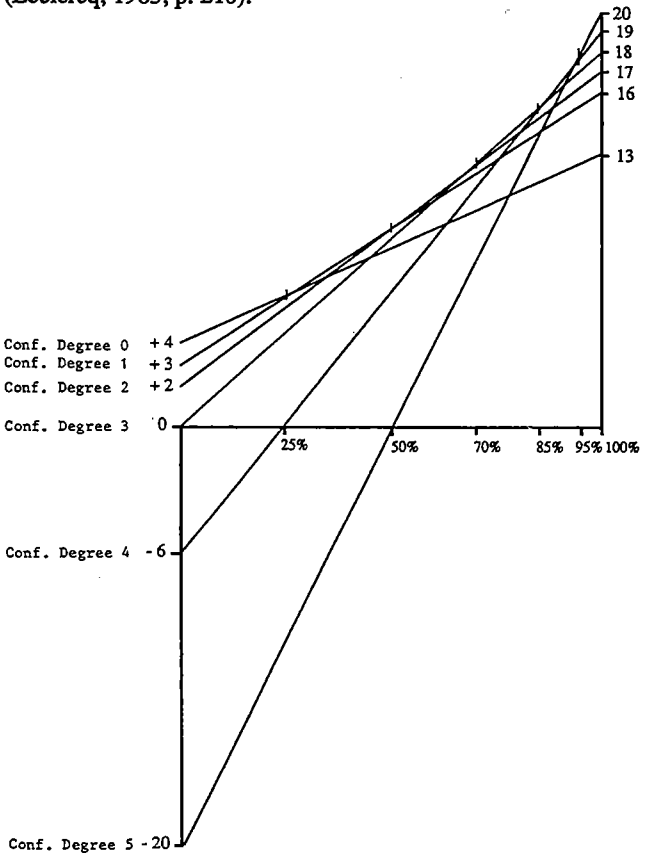


Figure 1

This graphic is the result of an operational research. It has to be "read" as follows. For each confidence degree, the "expected scores" (a theoretical concept) can be computed for any value of subjective probability, by the formula:

expected score (where using degree i with subjective probability being p)
$= (p.TC_i) + (q.TI_i)$

where TC_i is the Tariff (amount of points for i) in case of Correct answer
TI_i is the Tariff (amount of points for i) in case of Incorrect answer
(for instance, TC is +20 and TI is –20 for confidence degree 5).
p is the (subjective) probability of the learner (i.e. his/her confidence in his/her answer being correct).
q is the subjective probability of the learner's answer being incorrect.
Actually, q is 1 – p (if p is 0.8 then q is 0.2).

The expected score computed for all the values of p constitute a straight line. It can be drawn without computation since, for instance for confidence degree 5, the expected score is - 20 when p is 0 (the min.for p) and + 20 when p is 100 (the max for p). Joining those two extremes values creates the line of expected scores for each confidence degree.

In figure 1, the e.s. lines have been drawn for each of the six confidence degrees. It can be seen that there is only a definite sector where each e.s. line is optimal (i.e. overtops all the other e.s. lines), and this sector corresponds to what has been announced in the instructions, i.e. the limits (in probability terms) of confidence zones. This shows "optically" which confidence degree is *optimal* (i.e. gives the highest exptected score) for which probability.

Such a diagram has to be computed for each kind of segmentation of the probability scale. It can be drawn by trial and errors, but it may take time! A computer program has been written to insure this calculation (Leclercq, 1983, p. 200).

2.3 We must distinguish measurements from payments

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

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

Actually, the problem itself is incorrectly 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 the person is realistic!

2.4 Explicit feedback about realism must be provided

The CERT computer program computes several mathematical indices and outputs a graphic representation of realism. The following figures (2,3 and 4) present 3 examples of students' "realism" when using instructions hereover.

The student on the left (2) overestimates himself, the student of the right (4) underestimates and the student of the middle (3) is very well calibrated.

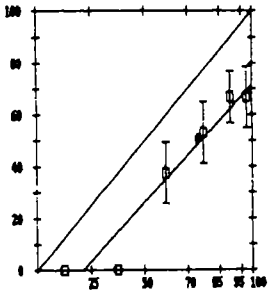


Figure 2

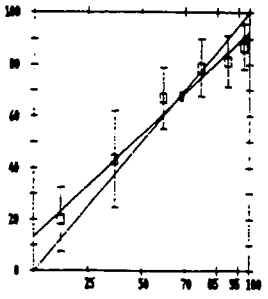


Figure 3

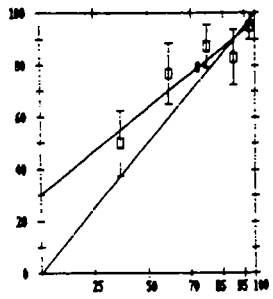


Figure 4

Each little "square" represents a rate of success for the given portion of the axis, i.e. for the five confidence zone. Each square is accompanied by the interval of confidence (standard error of measurement) of the percentage it represents (formula = 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 unrealistic.

2.5 Students must be trained

When they use confidence degrees for the first time, students usually adopt non optimal strategies (well known by decision making specialists, such as Lindley, Raiffa, Savage, Luce, Tversky, Kahneman, who have even given names to those strategies).

- choose the degree with the highest score (in case of success), i.e. an optimistic strategy (actually called *maximax*));.
- choose the confidence degree that insures the lowest "lost" (in case of failure), i.e. pessimistic strategy (actually called *maximin*);
- a mixture of the two (actually called Hurwicz's principle);
- always the same degree according to a fixed probability (for instance always degree 2, according to 60 % of chances) (actually called Atkinson's principle);
- give the probability corresponding to random guessing, i.e. the principle of total ignorance (actually called Laplace's principle);
- give the confidence degree that provides a given Variance (or difference) between the two tariffs, i.e. in case of success and in case of failure (actually called Coombs' principle);
- etc.

None of these strategies is optimal.

Neither verbal nor graphic nor equational explanations can convince the students that telling the truth is the best strategy in terms of maximizing their points. *They have to experience it.*

Only after the first trial, and above all the first *feedback*, do they admit a series of evidences: their strategy was bad, and their neighbours' strategies were even worse.

2.6 Adapt to situations

Sougne et al. (1990) suggested an original mode of answering to a multiple choice question :

- for each alternative (here 5), answer by T (true) or false (F);
- then, give *ONE confidence degree of this set of (5) answers* using the following probability scale (recommended by Leclercq, 1988).

Figure 5 illustrates this.

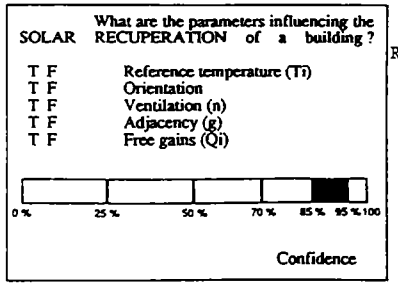


Figure 5

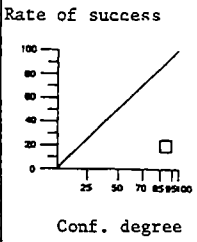


Figure 6

The advantage of such an approach is that an index (and a graphic of realism are available for each series of questions. Figure 6 illustrates the graphic for a person who answered 1 question correctly out of 5 (20 % correct) with a confidence degree of 90 %.

Figure 7 illustrates more classical alternative instructions and answer modes:

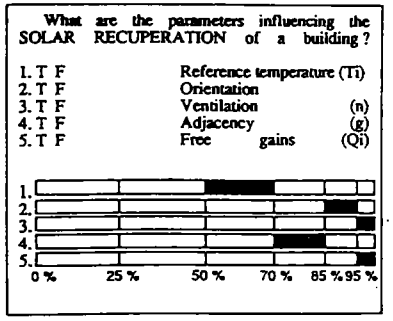


Figure 7

Of course only 20 groups of questions of this type provide 100 (T/F) answers and confidence degrees, enabling fruitful statistics.

The five principles described in this section have been followed in the experience that will be described:

- metric scale, asymetric to fit human sensitivity (acuity) and reliability
- tariffs computed according decision theory principles to reinforce the honest expression of probability estimates (i.e. telling the truth)
- distinguishing measurements (of competency and realism) from the payment of what the learner observes (actually a weighted combination of two measures).
- providing explicit feedback on realism with the help of a graphical representation (called the calibration graphic).
- familiarizing students with the system.
- adapt to situations: in some occasions, the learner will have to give one confidence degree for one single answer, in other occasions he/she will have to distribute his/her 100 % probabilities among alternatives, in other occasions, he/she will have to give only one probability value for a series of related answers.

The use of confidence degrees is not a necessary characteristic in the TASTE approach, but it has been incorporated in it from the beginning.

3 Open-book exams (the TASTE approach)

Open-book work is the most common situation in adult work, as well as in adult learning. Often, we discover new things to understand and memorize from written text. We read them, we try them (on the computer, the engine or the objects dealing with the content) and, *only if we do not understand*, do we ask others.

Since training can more and more rely on (well written and well illustrated) textbooks, videos and computer programs, it is sensible to apply within training the principles of adult learning conditions, in a formula called TASTE : "Towards an Adult System of Training and Evaluation".

The various steps of TASTE are as follows.

Step 1. *Before the course/exam day*

Learners (at home) read books, watch videos, etc.

Step 2. *On the course/exam day*

Learners ask the trainer questions about contents which were not understood. Those questions are mostly of two types.

The first ones adress prerequisite notions they are lacking. That is the bad news for the trainer (he/she is forced to re-teach unmastered contents), but for those learners it is of paramount importance that theses requests be met.

The second type of questions focus to the "top" of the content: ambiguities within one document, contradictions among various documents, connexions with quite different contents, future development, etc. This is the good news for the trainer because those interactions are rich, even for the teacher! He/she has sometimes the feeling of discussing with peers (colleagues), or at least with informed persons, what is scarcely the case in the classical lecturing situation.

Step 3. *The exam*

The trainer distributes the questions; learners answer on two sheets: one they give back to the trainer (who will introduce the data into the computer) and one

(identical) they keep for themselves. For each question, students are invited to provide three types of responses in a preprinted sheet : their answer, a confidence degree (see Leclercq, 1992) and justifying comments.

Q	A	C	Justification and comments
1			
2			
3			
4			
5			
Etc			

Q = Question A = Answer C = Confidence degree

Figure 8: Answer sheet for a student

Step 4. *Direct feedback*

The trainer/teacher communicates the correct answers and each student is free to oppose it, i.e. to argue in favour of other answers he/she considers should be accepted as correct. If the teacher is convinced of the correctness of the reasoning, this one is given more importance than the chosen coded answer, so that the teacher changes the given answer to the correct one ... provided the argument has been written in the justifying comments area of his/her answer sheet.

This step has immediate effects on the quality of questions (the teacher benefiting from remarks about each question), as well as the unambiguity of the reference document (it is sometimes questioned as well).

Some questions may reveal themselves to be so ambiguous that the teacher may immediately decide to suppress them so that the students' scores will not be affected by those wrongly phrased questions.

4 Interactive testing (The CHECK package)

CHECK is a series of interrelated softwares (mainly written in CLIPPER by H. Wuidar and partly in PASCAL by Ph. De Brogniez) that offer some evaluation facilities to the teachers and to the students.

4.1 CHECK IN

This software enables the teacher/trainer to add new questions (or modify existing questions) in one of his/her item banks.

Each question has the following structure :

- a code number;
- the phrasing of the objective (that will be printed on the students's report sheet);
- the phrasing of the question itself (with the suggested solutions);
- the instruction code (since several instructions concerning the type of question, the mode of answer and the tariffs are available);
- the correct answer code;

- a feedback (2 lines max.) that will be displayed on the screen just after the student has confirmed his/her answer and is told the correct one;
- a content code (optional);
- a mental process code (optional).

4.2 CHECK FROM

This software enables the teacher to select questions from the item bank to constitute (automatically) parallel versions either of paper-pencil tests or of an interactive test.

4.3 CHECK UP

This software enables each student to choose a given domain (within one of the various item banks) and to train (or practice) in answering a series of questions, being noted and receiving (immediate or delayed) feedback.

The successive steps of the test are as follows.

1° The student is told how many items are in the selected test.

2° The student is informed that
- all the questions are presented twice to him/her, but in a fixed order;
- during the first presentation, the student has to provide an answer and a confidence degree that is NOT scored;
- during the second presentation, the student has the possibility to change his/her previous answer and confidence degree, but they are now definitive and are not scored;
- the instructions, especially the codes for Reject (6), All (7), Lack of data (8) and Absurdity (9), are permanently displaied on the top of the screen;
- time spent is permanently displaied on the (top of the)screen;
- the clock is stopped between two questions (during this period, the students are invited to write their justification on a piece of paper).

3° The testing starts, item by item, for a first presentation where
- the question is displayed (and chrono starts);
- the learner answers (response + confidence degree), then is asked to revise or confirm.

4° The second presentation is the same except that, after each confirmation of the answer and the confidence degree, chrono is stopped and the learner is asked whether the correct answer and his/her scores should be communicated to him/her immediately (step 5) or at the very end of the test (step 7).

5° The correct answer, the feedback and the total score are displayed (if wanted). The student fills (handwriting) the justification area on his/her answer sheet.

6° Next question is displayed and chrono starts again.

7° When all questions have been answered, if wanted, all the feedback not yet communicated, are displayed one after each other.

8° At the end of the testing, a print out is produced with the help of CHECK OUT (see the following example). Nevertheless, the student is informed that the teacher will read his/her handwritten justifications and could change (favorably since justifications for correct answers will even not be read at all)

the report by introducing the correct answer for some incorrect responses and by producing a new report.

4.4 CHECK OUT

The software enables to print various types of reports. Here is an abbreviated version of such a report.

	FIRST RUN				SECOND RUN				
Quest.	Ans.	Conf	0/1	Time	Ans.	Conf	0/1	Time	Objective
1	3	5	0	136	3	2	0	127	Understand redacti
2									Understand what a
3									Detect redaction fl
4									Understand technic
5									Apply COOMBS, MI
6									Understand the imp
7									Apply SLAKTER's ap

Figure 9: Extract from a feedback sheet provided to a student after a CHECK sesssion

As can be seen, the student is not given the items themselves, to prevent students from learning the questions instead of the course.

5 Results

F. Lambert (1992), made the following observations on tests that were used for different chapters of two of Leclercq's courses (course 155 with 20 students, course 58 with 120 students). Suppressed questions are those that the teacher decided, either on the basis of the students answers or comments, or on the basis of item analysis, to suppress after the students had answered.

5.1 Mean time to answer and status of the questions

Type of questions	C155 Test 1 20 students 12 questions	C155 Test 2 20 students 15 questions	C155 Test 3 20 students 12 questions
Correctly answered	162	245	205
Incorrectly answered	215	277	276
Suppressed	262	295	307

Figure 10: Mean time in seconds for answering

Repeatedly, in this course (on evaluation) correctly answered questions need less time than incorrectly answered ones and far less than suppressed ones. So testing time is reduced by the students' abilities (less incorrect answers) and by the questions quality (less suppressed questions).

5.2 Time to answer and type of question

The following observations are averaged from test 1 (122 students of second year of graduate studies in psychology answering 12 questions) and test 2 (120 out the same students answering 12 other questions) of course 58 (educational technology).

Out of those 2904 answers, only 2134 (73%) have been considered because they have not been modified after the teacher has read the students' justifications.

Average data for the 4 types of general implicit solutions (GISs) are as follows :

	Number of answers	Time in seconds	Rate of success
Usual	708	269	68 %
Reject (6)	399	297	59 %
Total (7)	406	226	66 %
Missing (8)	245	229	57 %
Absurdity (9)	376	179	74 %

Figure 11: Statistics concerning 5 types of solutions

It appears that the *Absurdity* solution is the quickest and the most correctly answered. *Lack* of data has the lowest rate of success and *Reject* takes the longest time to answer.

5.3 Discrimination indices

The previous results deserve to be compared to the observations made in may 1987 on 140 students (fourth year of graduate studies, from various orientations : sciences, linguistic, social sciences, economics, etc.) having answered 82 MCQs with 4 kinds of SGIs :

Correlation	Number of quest.	Aver. success	Aver. point biserial
Usual	40	66 %	0,29
Reject	9	48 %	0,38
Total	11	55 %	0,31
Missing	10	28 %	0,29
Absurdity	10	66 %	0,42

Figure 12: Average characteristic of 5 types of items

Here again, the most difficult appeared to be the *"lack of data"* type of question and the absurdity the most frequently correct (with the usual type of questions).

About the same order in average difficulty is observed in the two experiments.

ABSURDITY and REJECT solutions have the best point biserial correlation indices.

It is worth nothing that *the average discrimination indices of GISs are higher* than the usual questions ones!

5.4 With and without book

As a part of the 1988 experiment, 22 questions had to be answered twice : the first time without the book and the second time with the book. There were four questions of each of the five types of questions, plus two fictitious questions. Those fictitious questions concern inexisting contents, according to Slakter, Koehler and Hampton 's idea (1970). We considered, as correct answers for those questions (non announced in the instructions) three answers : "Omit", "Lack of data" and "Absurdity".

Here are the results (on 140 students) :

	Main facilities			
Categories	Number of questions	without the book	with the book	GAIN
Fictitious	2	47 %	62 %	+15 %
Usual	4	50 %	54 %	+ 4 %
Reject	4	40 %	48 %	+ 8 %
Total	4	55 %	60 %	+ 5 %
Missing	4	33 %	35 %	+ 2 %
Absurdity	4	58 %	65 %	+ 7 %

Figure 13: Average characteristics of 6 types of items

As foreseen, the *greatest gain* happens for the *fictitious* questions (the book had an index of keywords and an other index of authors names).

After that, *REJECT and ABSURDITY benefited the most from the book*, and LACK benefited the least.

Gains are deceptive in average. This can be partly explained by the fact that maybe students have not be given enough time (browsing into a book requests a great amount of time) and by the fact that most of those students were not trained to this kind of answering (to those instructions).

6 The students' opinion

F. Lambert (1992) asked to the 140 students (second year of graduate studies in psychology and education) of course 58 (educational technology) to answer an opinion questionnaire. Here are the results to various questions:

6.1 The FAIREST type of questions (N = 119)

Students had to grade each question on a 10 degrees scale about which system is the fairest. Average results are as follows:

MCQ with GIS and open books and justification comments	6,53
Written exams with long answers	5,22
Oral exams	5,15
Written exam with short answer	5,22
Classical MCQs	4,31

Figure 14: Students' opinions about several types of questionning methods

F. Lambert has been very cautious to avoid the "desirability phenomenon" (written commitment not to reveal to the professor who has answered what). Nevertheless, this phenomenon is still likely to have occurred. The following results to other questions could nevertheless help understand the causes for the preferences.

6.2 Appreciations of justifications (N=124)

Students were asked:

a) whether they *strongly dislike* (1) or *strongly like* (7) (on a 7 positions scale) the principle of their writing justifications, these ones been read by the teacher only in case of incorrect answer and constituting a second chance for being credited with a correct answer;

b) whether (7) or not (1) the possibility of winning points by written justification compensates time it needs to write them down.

Results are as follows:

	a		b	
I strongly dislike	1	4.0	Does not compensate	6.6
	2	7.3		9.0
	3	4.0		7.4
	4	8.1		13.1
	5	12.9		13.9
I strongly like	6	21.8		21.3
	7	41.9	Compensates	28.7
		100.0		100.0

Figure 15: Students' opinions about the justification principle

Subhypothesis like "the students who actually use *and benefit* from this piece of instruction are more favorable than others" have not been tested yet (whereas data would permit that).

6.3 Preferences for oral or written exams (N=121)

Those two possibilities were the opposite "poles" of a seven degrees scales.

oral	1	9,9 %
	2	14,9 %
	3	13,2 %
	4	21,5 %
	5	20,7 %
	6	14,0 %
written	7	5,8 %

Opinions vary broadly, no consensus is possible. This is very important for examiners to know that *probably there does not exist a single instruction that will fit all students !*

Therefore, one should consider instructions that can be adapted to the learner's preferences on his/her own request.

6.4 Preferences for open (long) answer questions or MCQs (N=121)

Open long answers	1	3,3 %
	2	13,2 %
	3	5,0 %
	4	11,6 %
	5	18,2 %
	6	27,3 %
MCQs	7	21,5 %

Whereas there is a tendency of a greater appeal of MCQs, preferences are spread along the whole continuum.

6.5 Preferences for Open (short answer) questions or MCQs (N=120).

Open short answers	1	5,8 %
	2	11,7 %
	3	12,5 %
	4	22,5 %
	5	13,3 %
	6	22,5 %
MCQs	7	11,7 %

Opinions are largely spread on the continuum and the advantage for MCQs is lower than it was for long answers to open questions.

6.6 Among MCQs, preferences for GISs or classical (N=122)

Classical MCQs	1	45,9 %
	2	18,9 %
	3	11,5 %
	4	5,7 %
	5	3,3 %
	6	7,4 %
GIS MCQs	7	7,4 %

Preferences are largely in favor of classical MCQs. Note that those students have never been shown statistical data about rates of success of the five types of questions.

6.7 How far is each of the SGIs appreciated?

	None reject	All of them	Lack of data	Absurdity
	6	7	8	9
1. strongly dislike	12,1	8,4	44,5	34,2
2.	6,5	4,2	18,5	13,3
3.	14,5	14,3	11,8	10,8
4.	31,5	31,9	10,9	13,3
5.	25,8	26,1	5,0	15,8
6.	7,3	10,1	6,7	5,0
7.strongly like	2,4	5,0	2,5	7,5
Total	100	100	100	100
Average	3,84	4,13	2,44	3,08
N	124	119	119	120

Figure 16: Students' opinion upon 4 types of General Implicit Solutions

The reluctance for LACK of data is in accordance with the low rate of success for this kind of questions. On the opposite, the (lowest) reluctance for ABSURDITY contrasts with the fact (insufficiently known by the learners) that this kind of question is the easiest one, i.e. elicits the highest rate of success.

6.8 Preferences for paper or computer administered MCQ SGI tests (N=119)

Computer administered	1	31,1 %
	2	11,8 %
	3	5,0 %
	4	13,4 %
	5	9,2 %
	6	13,4 %
Paper administered	7	16,0 %

Again, responses are spread, with an advantage to computer adminstered test.

6.9 Comments on advantages and drawbacks of MCQs GISs

The end of the questionnaire asked for the major advantages and drawbacks of
- MCQ GISs with open books followed by collective correction;
- Computer administered testing.

1° MCQs GISs with open books:

Main drawbacks	Main advantages
Ambiguous questions forcing to look systematically for traps and implying a perfect understanding of written language. (25 %)	Addresses reflexion and comprehension instead of rote learning, and leads to a better mastery of the content. (32 %)
Increased attention, concentration, stress, tiredness. (20 %)	Immediate feedback : score is known the same day. (8 %)
Confusion among GISs : 6-7-8-9 ??? (20 %)	The type of requested answer is clearly specified and the student is not anxious to have "enough" answered. (5 %)
Others : less than 5 %.	Others : less than 5 %

2° Computer administered testing

Main drawbacks	Main advantages
Impossibility to come back to a previously answered question, no overview of the whole set of questions. (33 %)	Immediate reception of feedback about correctness of responses. (44 %)
Being informed immediately of one's errors and lost of points has a negative effect on motivation. (14 %)	It frees successful students from being examined an other time on the same content. (11 %)
Fear of key stroke failures, computer breakdown, etc. (10 %)	Allows to practice, to see whether the way of studying is correct. (6 %)
Increased attention, concentration, stress. (7 %)	Allows justifications. (6 %)
	Lowers the fear of failure in front of a human person. (6 %)
	Introduces variety, ludic behaviour, change in learning. (6 %)
	Saves time (important in exam period). (5 %)

The interviewed students have been presented the questions only once. This explains the main drawback (impossibility to come back). A TWO RUNS version has been introduced afterwards, on the basis of their opinion.

The possibility of not being presented directly one's score (and, of course, the correct answer) has also been introduced as a consequence of the second main drawback (negative effect on motivation).

The actual decisions made by the testees are now recorded (time spent on each question during the first and the second run, same for answers and confidence degrees) and analysed (stability in correct answer; good move, i.e. change an incorrect answer into a correct one; bad move, i.e. the contrary; stability in error; changing an error for an other one, etc.). We hope this will enable us to detect cognitive strategies and to provide useful feedback on metacognitive processes.

7 The teacher's role and opinion

7.1 The teacher has to spend time on:

a) Defining objectives (they can actually be contents: sections of a chapter, concepts and subconcepts, etc.)
b) Designing MCQs (with GIS) within these categories of contents and with an appropriate feedback that will appear on the screen.
c) Reading learners' written comments accompanying their answers (NB: when the answer is correct, it has been announced to the students that the learner's (handwritten) comment will not be read at all by the teacher so that this process can only be beneficial to the student) and change the student's answer to the correct one (if comments are convincing).
d) Change the item so that the same confusion cannot happen further.
e) When d (changes in students' responses) is finished, introduce (through keyboard or optical reader) the answers within a computer program (CERT).
f) Analyse item characteristics such as rate of selection of the various alternatives, their discrimination index and (on the basis of this last index), discard some items.
g) Compute the students' score only on the remaining (validated) items.

7.2 Difficulties and advantages

Operations a and b described above are time consuming, but they result in a highly coherent course in which students know in advance very well what the criteria will be (the objectives are given to them) and in which they receive a detailed feedback from assessment procedures, facilitating the remediation process. Operation d is time consuming too, but it results in an item bank constantly increasing in quality (teachers need cumulating their expertise) on experimental grounds.

Operation c and f can make difficulties for teachers to accept that the quality of their questions can be disputed, but students dislike even more teachers who do not accept any remark! Which teacher can insure that his/her questions are always perfect?

Conscious of the intent of fairness and of the teacher's good will, the students have also the feeling of participating in the process, i.e. that their comments are actually taken into consideration and can have an influence on the system itself.

7.3 New adepts

The TASTE CHECK approach has recently been used (November 92 to January 93) with the (250) graduate students of the University of Liège who prepare themselves in becoming secondary highschool teachers. They followed Leclercq's course on Educational Psychology and were originated from all faculties (from romanists to veterinarians). Their advices are being collected.

Professor Born, Leclercq's colleague at the University of Liège, has engaged the whole process for his course about developmental psychology for (200) freshmen students in psychology.

A.C. Nizet, who assisted Prof. Born, has proposed intelligence and cognitive style tests to these students. This research is in development. A key concern of Prof. Born was that all the students will anyway be assessed in the oral mode, but the interview of excellent students (on the basis of their TASTE scores) can be facilitated (and accelerated).

8 General discussion

The computer administration of GIS MCQs (here with CHECK) presents problems. One of them is the duration of testing. Usually, we allow 90 min for 15 questions, including the time needed by the student to read the feedback displayed by CHECK after each answer. Therefore, some students have not time to see all the questions two times and they are therefore scored on the basis of their first answers to some questions.

General Implicit Solutions (GIS) are only a first step in a more sophisticated way of questioning. In a next version, students who indicate "Absurdity" will be asked to point (with the mouse, or in a written way) where the absurdity is. The same for the answer "Missing - lack of data": they will be asked to specify which kind of information they would like to obtain; they will be given it and invited to answer a second time. The same for the "Reject-None" answer: the student will be invited to type *his/her* correct answer.

As can be seen, the barriers between multiple choice and open ended questions are slightly disappearing, as well as the barriers between training (practicing) and evaluation (scoring).

The student will be more and more placed in a professional situation, i.e. to process "cases". He/she will be more and more in the same position as a physician who receives a series of individual cases in his consultation, detecting "usual" situations, discarding "absurd" diagnosis (often coming from the patients themselves), obtaining more information (inspection of the body, analysis of blood or urine, x-rays) in case of lack of data, detecting multiple disease situations, etc.

It is well known that medical experience comes from not only the exposition (in a passive way) to numerous cases, but essentially from the actual confrontation with cases in a problem-solving situtation.

MCQs with SGI, confidence degrees, the CHECK interactive approach and the TASTE open books principles are steps into this direction: learning by living,

even if it is in a simulated way, since education is bound to accelerate the process of learning. How fast it can be accelerated not only on specific contents, but on the enquiry process.
The answer to this question is not around the corner, but seems now less unreachable. We even have the conviction that the whole approach described here is a good direction to be followed and deepened to approach this challenging goal.

References

Barras, H. (1992), QCM sur Minitel en EAD: vers l'automatisation, in Weber & Dumont "Les questionnaires automatisables", Marne-la-Vallée: Colloque de l'ESIEE

Boxus, E. (1988a), Les QCM à solutions générales au service de l'évaluation à livre ouvert, in *Actes du Colloque International "Formation, Evaluation, Sélection par Questionnaires Fermés"*, Marne-la-Vallée: ESIEE, vol. 1, 318-331

Boxus, E. (1988b), Vers des examens à livre(s) ouvert(s)? In *Le défi pédagogique de l'enseignement supérieur, Actes du Congrés de l'Association Internationale de Pédagogie Universitaire,* Montréal: AIPU, 533-537

Boxus, E. (1992), Check: une banque de questions interactives, in Weber & Dumont: questionnements automatisables, Colloque International ESIEE: Marne-la-Vallée, 29-49

Boxus, E., Leclercq, D. Osterrieth, S., Wuidar, H. (1991), Principes communs pour évaluer les résultats cognitifs de la formation, Bruxelles: Commission des Communautes européennes, EUROTECNET

Bruno, J. & Baxter, J. (1989), An application of information reference testing. Proceedings of the Sixth International Conference on Technology and Education, Orlando, vol. 2, 191-192

Bruno, J.E. (1987), Using MCW-APM test scoring to evaluate economics curricula. Journal of Computer Boxed Instruction

Bruno J.E. (1990), Confidence contour tests item analysis with information referenced testing. Paper presented at the Seventh International Conference on Technology and Education, Brussels

Bruno, J.E., Baker, J.B. (1989), Computer assisted learning through technology based classroom formative evaluation: an application of information referenced testing. Proceedings of the Sixth International Conference on Technology and Education, vol. 2, Orlando, Florida

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

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. J. of Mathem. and Stat. Psych. 18, 87-123

De Landsheere, G. (1984), *Evaluation continue et examens, Précis de Docimologie,* Bruxelles: Labor

Depover, C. (1987), *L'ordinateur média d'enseignement,* Bruxelles: De Boeck

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

Dressel, P. & Schmid, J. (1953), Some modifications of the multiple-choice items. Educational and Psychological Measurement 13, 574-595

Dudley, H.A.F. (1973), Multiple-choice tests. Lancet 2, 195

Dudycha, A.L. & Carpenter, J.B. (1973), Effects of item format on item discrimination and difficulty. J. Application Psychology 58, 11-121

Dumont, B., Lazerges, J.M. (1992), Des QCM télématiques en mathématiques pour préparer un concours d'entrée dans une école d'ingénieur, in Weber & Dumont, "Les questionnaires Automatisables", Marne-la-Vallée : Colloque de l'ESIEE

Fabre, J.M. (1975), *Docimologie expérimentale et évaluation par questionnaires : étude du jugement multiple et de l'autopondération,* Thèse de doctorat de troisième cycle en psychologie, Université de Provence, ronéotypé

Fabre, J.M. et Noizet, G. (1977a), Confiance attachée aux réponses à des questions à choix multiple, *Journal de Psychologie Normale et Pathologique,* 74, 335-362

Halleux-Hendrick, J. (1969b), Construction des questions à choix multiple: une seule solution correcte?, in *Revue Belge de Psychologie et de Pédagogie,* Tome XXXI, n°127, 113-125, Bruxelles

Hughes, H. & Trimble, E. (1965), The use of complex alternatives in multiple-choice items. Education and Psychology Measurement 25, 1

Hunt, D.P. (1993), Human self-assessment - theory and application to learning and testing. In Leclercq & Bruno (eds.), Item Banking: Interactive Testing and Self-Assessment. NATO ASI Series F, Vol. 112. Berlin: Springer-Verlag (this volume)

Kagan, J., Impulsive and reflective children: significance of conceptual tempo. In J. Krumboltz (ed.), Learning and the Educational Process. Chicago: Rand McNally, 133-161

Karraker, R.J. (1967), Knowledge of results and incorrect recall of plausible multiplechoice alternatives. J. Education Psychology 58, 11-14

Keller, F. & Sherman, G. (1974), The Keller Plan Handbook. Menlo Park, CA: W.H. Benjamin

Keller, F. (1968), Goodbye teacher. J. Applied Behaviour Analysis 1, 78-89

Lambert, F. (1992), Evaluation interactive informatisée et auto-évaluation, graduation Dissertation, Faculty of Psychology and Education, University of Liège

Leclercq, D. (1980), Computerised tailored testing: structured and calibrated item banks for summative and formation evaluation. European Journal of Education 15(3), 521-260

Leclercq, D. (1983), Confidence marking, its use in testing. In B. Choppin and N. Postlethwaite (eds.), Evaluation Education, an International Review Series, 6(2), 161-287n

Leclercq, D. (1986), *La conception des Questions à Choix Multiple,* Bruxelles: Labor.

Leclercq, D. (1987), *Qualité des questions et signification des scores,* Bruxelles: Labor

Leclercq, D. (mars 1988), Mesurer la connaissance partielle et le réalisme par les degré de certitude, in Actes du Colloque International "Formation, Evaluation, Sélection par Questionnaires Fermés", Marne-la-Vallée, vol. 1, 306-316

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

Lindley, D. (1971), Making Decisions. London: Wiley

Luce, R. & Raiffa, H. (1966), Games and Decision. New York: Wiley

Luce, R. (1959), Individual Choice Behavior. New York: Wiley

Massengill, H. & Shufford, E. (1967), What pupils and teachers should know about guessing. Technical Report SMC R-7, Lexington, MA

Noizet, G. & Caverni, J.P. (1978), *Psychologie de l'évaluation scolaire,* Paris: PUF

Noizet, G. & Fabre, J.M. (1975), Etude docimologique des questionnaires à choix multiple (QCM) : Perspectives de recherche, *Scientia Peadagogica Experimentalis,* vol. 12, 38-62

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

Pyrczak, F. (1972), Objective evaluation of the quality of multiple-choice test items designed to measure comprehension of reading passages. Read. Res. Quart. 8, 62-71

Pyrczak, F. (1974), Passage-dependance of items designed to measure the ability to identify the main ideas of paragraphs: Implications for validity. Educational and Psychological Measurement 34, 343-348

Savage, J. (1971), Elicitation of personal probabilities and expectations. J. American Statistical Association 66, 336, 783-801

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

Shuford, E. (1993), In pursuit of the fallacy: resurrecting the penalty. In Leclercq & Bruno (eds.), Item Banking: Interactive Testing and Self-Assessment. NATO ASI Series F, Vol. 112. Berlin: Springer-Verlag (this volume)

Slakter, M., Koehler, R. & Hampton, S. (1970), Learning test-wiseness by programmed texts. J. Educational Measurement 7, 247-254

Smith, R. (1970), An empirical investigation of complexity and process in multiple-choice items. J. Educational Measurement 7, 33-41

Strang, H.R. & Rust, J.O. (1973), The effects of immediate knowledge of results and task definition on multiple-choice answering. J. Experimental Education 42, 77-80

Strang, H.R. (1977), The effect of technical and unfamiliar options on guessing on multiple-choice test items. J. Educational Measurement 14, 253-259

Tversky, A. & Kahneman, D. (1974), Judgment under uncertainty: Heuristics and biases. Science 185, 1124-1131

Von Neumann, J. & Morgenstern, O. (1974), Theory of games and economic behaviour. Princeton University Press

Wahlstrom, M. & Boersma, F. (1968), The influence of test-wiseness upon achievement. Educational and Psychological Measurement 28, 413-420

Williamson, M.L. & Hopkins, K.D. (1967), The use of "none of these" vs homogeneous alternatives on multiple-choice tests: Experimental reliability and validity comparisons. J. Educational Measurement 4, 53-58

Wood, R. (1974), Guessing on objective type test items. School Science 56, 179-180

Wood, R. (1976), Inhibiting blind guessing: the effect of instructions. J. Educational Measurement 13, 297-307

Wood, R. (1977), Multiple-choice: A state of the art report. In Choppin & Postlethwaite (eds.), Evaluation in Education International Progress. Oxford: Pergamon