European Journal of Education, Vol. 15, No. 3, 1980

Computerised Tailored Testing: structured and calibrated item banks for summative and formative evaluation

DIEUDONNE LECLERCQ

Educational Evaluation in Progress

An educational evaluation is always part of a regulation process, and regulation processes in education have already been fully discussed elsewhere (D. Leclercq, 1976, 1978a and 1978b). Evaluation serves objectives that can be obvious, or that can be less evident, such as abilities certification (and thus *selection*), achievement prediction (and thus *orientation*), and individual proceedings analysis (*formative purpose*).

The first two functions of evaluation (selection and orientation) have been developed in education since the beginning of this century in the field of psychometrics. The key-concepts are those of traditional test theory, *i.e.*, test standardization, validity, reliability, and accuracy of scores. Psychometrics corresponds to summative evaluation, that seeks one stable and representative score.

The third function of evaluation (formation) seeks analytical and causal information. The *mastery learning* approach has been settled on the principles of early diagnosis and immediate remedial treatment. Behaviourists, such as B. F. Skinner, have shown that a frequent and quick feedback can have a positive effect on the acquisition of abilities. Mastery learning illustrates the educationists' awareness of the dangerous 'normal curve myth', so that 'edumetrics' can now be developed.

More recently, other methods have appeared, enabling self-regulation (self-diagnosis, self-correction, *etc.*) and self-assessment (in which the student expresses, by means of subjective probabilities, how confident he is of the correctness of his answer). A number of theoretical and practical works dealing with this problem have been published (D. Leclercq, 1975 and 1977a, b and c). Not only do these methods open up new prospects for educational assessment, but they also offer the advantage of placing the student at the centre of his own learning, of having him assume the responsibility for his competence, and of getting him involved in his own development.

Meanwhile, psychometrics has also evolved. The traditional test theory has been thoroughly revised in the light of the 'generalisability theory' (Cronbach *et al.*, 1972, and Tourneur & Cardinet, 1978). The concepts of individual ability (A) and item difficulty (D) have been re-examined in the light of the Rasch model, in which the probability of a right answer to an item depends only on the individual's ability (A) and on the difficulty of the item (D). This probability is noted P(1 | A, D), where P stands for 'probability', 1 for 'success', and the sign | means 'given'. The two parameters (A) and (D) are expressed in the same units and plotted on the same (horizontal) scale of measurement.

As a unit, B. Choppin (1978) proposes a WIT scale, wherein A and D values vary,

in practice, from 20 to 80. The word WIT has been coined from the word BIT, the letter W standing for the number 1·24573, which offers interesting numerical properties, among which are that $W^{10}=9$, $W^{5}=3$, $W^{-10}=\frac{1}{9}$, and $W^{-5}=\frac{1}{3}$. In Choppin's formulation of the Rasch Model, the basic equation is the following:

$$P \mid (iA,D)=\frac{W^{A-D}}{1+W^{A-D}}$$

This model constitutes the basic reference for *calibrated* item banks where each question has its own difficulty index (in WITS), and not simply a p-index measuring the percentage of success observed on a given sample of students, because the p-indexes depend upon the students' abilities. In fact, the same question will be given high or low p-values, depending on whether the students are more or less proficient. Consequently, a series of p-values is needed for population ranked according to growing ability. All those p-values plotted on a scale would reveal a *characteristic item curve.* An important point in the Rasch Model is that the difficulty of an item is not expressed by a single number, but by a curve, a function. This happens to be a logistic function since the logit formula $\left(\frac{a}{1+a}\right)$ appears in the basic equation.

Several variants of the Rasch Model are currently being developed. In German-speaking countries, G. Fischer (1975 and 1977) from Vienna and H. Spada (1977) from Kiel put forward various programmes in order to estimate A and D values. In the United States, Lord (1970) and Lord & Novick (1968) developed, from the so-called Birnbaum Model, the 'latent trait theory', wherein an item is described not only by its D (difficulty) parameter, but also by two additional parameters (discrimination and 'floor' value).

An immediate application of such a model is *tailored testing.* For example, knowing the student's ability (A), the computer presents a question with such a D(difficulty)-index that the probability (P) of success is 0·5. It has been shown, (Wood, 1975) that the information derived from a success or from a failure is maximal for an *a priori* probability of success of 0·5. This process enables psychologists to estimate more quickly and with fewer questions the individual's true score or ability.

Three Types of Item Banks

Item Pools

Since the sixties, such item pools have broadly developed in the United States and in Europe. Contrasting with tests, the stored items are here totally independent of one another. By way of illustration, I shall describe the Belgian Air Force Item Pool which was set up in 1971 with P. Van Roy and others.

Multiple choice questions are card-indexed. The questions are typed in French on one side of the cards, and in Dutch on the other. They are classified according to a code number referring to the content. Psychometrical features of each question (objective popularity, subjective attractiveness, and discrimination indices of each of the alternatives) have been previously stored in various ways. At first, the sheet of paper containing the indices was stuck on the item card, but this listing was soon replaced by a punch-card, which made it possible to get immediate computer comparisons between present and past indices. At present, these data are collected on direct access disc files.

By the end of the first year, the number of questions reached 300, and eight years later 20,000. Improvements, by now, are mainly aimed at quality.

Situated in an Evaluation Service Centre, the bank helps nine schools in building and correcting tests as follows. Teachers first submit their requirements regarding tests (such as those concerning choice and order of the questions). The Centre then builds

the tests and provides the number of copies wanted, as well as special answer grids (on self-copying paper). These strips of paper are sent back to the Centre where the students' answers are punched on cards and corrected by a FORTRAN computer programme, (BANKETFA). A very comprehensive PL1 programme called STEP is to be found in Debot & Leclercq (1978).

Although the independence of the questions enables the teacher to conceive any possible kind of test, he still has to work a great deal if he wants a structured test.

Since the questions do not refer, in this item pool, to behavioural objectives, a number of problems arise in regard to classification, copying of already existing material, and interpretation of the results. A taxonomical index has been given to each question (Bloom's taxonomy reduced to four categories), but ambiguity is still not completely eliminated since the taxonomical level depends on the students' previous experience: creativity can turn into rote memory if the student has already encountered the same difficulty. Moreover, studies on the Rasch Model show how ambiguous item difficulty indices can be, while repeatedly building new paper tests is an expensive solution.

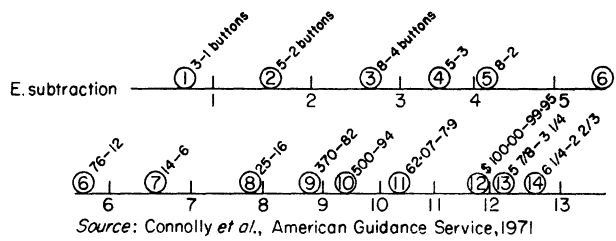
Despite these weaknesses, this item pool, together with the evaluation service, have proved highly efficient in the sense that they have made it possible to process a considerable quantity of data in a very short time, and to carry out original experiments, especially those requiring follow-up work and repeated testing.

The Evaluation Service Centre now constitutes a permanent research department and, together with the University of Liège, has greatly helped other educational establishments to develop their own item pool systems.

Calibrated Item Banks

In a calibrated item bank, each item (or question) is associated with its 'item characteristic curve' or ICC. B. Choppin & N. Postlethwaite have already made a substantial contribution to the conception of such a bank in Indonesia (Nasoetion *et al*, 1976). In their model, all the curves have the same slope, differing only in position (the more to the right the ICC, the more difficult the item). Since the centre of the curve is always at the 0·5 probability level on the vertical axis, an ICC can be described by the sole co-ordinate of the centre on the horizontal (difficulty) axis. Questions can then be placed on a unique continuum according to this index. Distances in difficulty thus appear between the items.

The example shown in Fig. 1 is extracted from the 'Keymath Diagnostic Profile' or KDP.



Source: Connolly *et al.*, American Guidance Service, 1971

FIG. 1.

The numbers beneath the axis represent the difficulty scale chosen for the KDP (this scale is different from the WIT scale used by Choppin & Postlethwaite, and it can be seen that the D-index of question 6 varies from 5 to 6. The fourteen questions scaled are shown in Table I:

TABLE I

(4)	(5)	(6)	(7)	(8)	(9)
5	8	76	14	25	370
−3	−2	−12	− 6	−16	− 82
(10)	(11)		(12)	(13)	(14)
500	62·07		$100·00	$5\frac{7}{8}$	$6\frac{1}{4}$
− 94	− 7·9		− 99·95	$-3\frac{1}{4}$	$-2\frac{2}{3}$

It should be noted that the KDP offers different scales for numeration, fractions, geometry, *etc.*, and that the D-index is in fact the ICC inflexion point, since logistic curves are S-shaped.

Fig. 2 indicates five ICCs on a WIT scale: their inflexion points are 50, 55, 60, 65 and 70 WITS respectively (abscisses of V, W, X, Y and Z points).

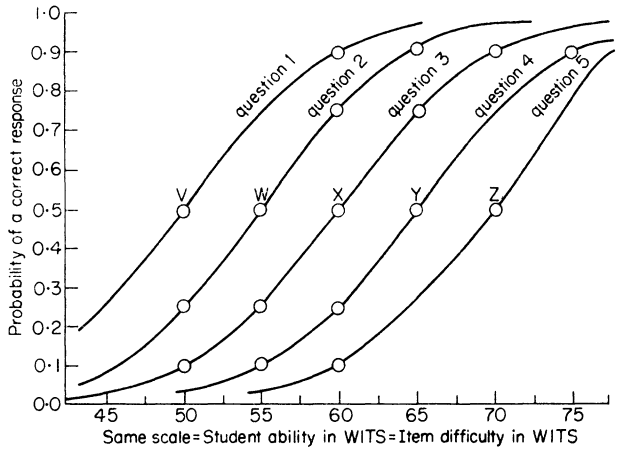


FIG. 2.

If a student of ability 60 WITS (A=60) is faced with question 3 of difficulty 60 WITS (D=60), the probability of success will be 0·5 (if A=D, then P=0·5) as shown by the point marked X. If the same student (A=60) is given question 1 (D=50), the probability of success may be calculated as follows:

$$P=\frac{W^{A-D}}{1+W^{A-D}}=\frac{W^{60-50}}{1+W^{60-50}}=\frac{W^{10}}{1+W^{10}}$$

giving a probability of success of 0·9. If the student gets question 2 (D=55), the corresponding calculation is:

$$P=\frac{W^{60-55}}{1+W^{60-55}}=\frac{W^{5}}{1+W^{5}}$$

or a probability of success of 0·75. Faced with question 4 (D=65), he will have 0·25 probability of success, while with question 5 (D=70), he will only have 0·10 probability of success.

Student A-values and item D-values are computed from the raw data matrix of successes and failures for all students to all items (Wright & Stone, 1979). When the

item D-values are known, it is possible to match every total score (except when the latter is 0 or the maximum) to an A-value on the WIT scale. In tailored testing, this is how the student's ability is *first* estimated. *Secondly,* the person conducting the experiment (or more frequently the computer programme) selects an item with a D-index equivalent to the A-value, so that the probability of success is 0·5.

R. Wood (1975) and others have developed such interactive tailored testing on the basis of the first two steps described above. The failure or the success of the next question leads to a new estimation of A (*third* step). In this 'up and down method', the difficulty of the questions varies as shown in Fig. 3:

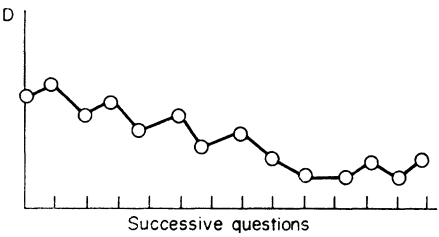


FIG. 3.

The same figure also shows the evolution of question D-values for a student who has been over-estimated at the initial stage. Wood (1975) considers that 80% of the measurement error variation is reduced after 15 questions, and 85% after 20 questions. He stresses the fact that this process is time-saving since it discards items that would undoubtedly be either failed or passed.

Structured Item Banks

In *La fonction régulatrice de l'évaluation vue sous l'angle de l'implication de l'étudiant* ("The regulative function of evaluation with regard to the student's involvement"), I emphasized the point that, in most educational assessment processes, the student is reduced to a mere executor: he just has to provide answers to questions he receives. I also suggested that the student himself should be allowed to:

(i) choose the objectives on which he is going to be evaluated (he would then become the *initiator* of the regulation process);
(ii) perform the measurement, by correcting his own answers and by scoring his own performances (he would then become the *observer*);
(iii) compare the results obtained to the initial goals, interpret the situation and determine the consequent strategy (he would then become the *decision-maker*).

These procedures met with two serious difficulties: the first is psychological, while the second is technical. Let us first turn to the psychological hindrance. Students could be suspected of choosing objectives which were too easy, cheating in the process of correcting and scoring, and interpreting the results so as to follow easier pathways in the future.

However, it is possible to counter these pessimistic yet fully understandable fears. In current practice, educational objectives are now commonly considered as an agreement to be concluded between the teacher and the students. These contracts entrust the students with an important part of the responsibility for the attainment of the objectives. With this in mind, the student assesses his abilities when he feels ready and in the fields he likes best, but in regard to objectives to which he has committed himself since the

beginning of the year. Y. Tourneur (1975) has shown how efficient it proved, from the standpoint of scholastic achievement, to convey the objectives to the students. The teacher will inform the students that the available assessment techniques are opportunities, and not compulsory methods, and will go on developing his own evaluation instruments.

The problem of cheating appears to be the same as in the case of programmed learning a few years ago. It was feared that the student would read the correct answers before giving his own. Yet teachers who have been using programmed courses will testify that this danger does not exist when the activity has been properly motivated.

Involving the students in the assessment process meets with a technical difficulty: as a matter of fact, one must have readily available tests referring to clearly stated objectives, correction criteria, remediative comments, and so on or, in other words, *self-assessment modules* (SAMs).

The main purpose of such SAMs is to train the students in assessing themselves, in order to train them how to learn. The five components of a SAM are:

(i) the objectives (so that the student can decide whether to use the SAM or not);
(ii) an answer sheet (that will help the student in the process of correcting his answers);
(iii) a set of questions;
(iv) correction rules (including the correct answers, comments about mistakes, and scoring formulae);
(v) a body of advice (to be given according to the various performances).

A split presentation of the questions facilitates a taxonomical approach. The SAMs are divided into three parts:

(a) *knowledge (memory) items* (which are corrected before going further into the SAM);
(b) *comprehension and application items* (which are also corrected before going any further);
(c) *higher cognitive processes items* (including complex applications, deep analysis, creativity, expression, *etc.*).

To prevent the students from modifying previous answers, the teacher may collect the answers at each step. With this split dispensation of items, failures at stages (b) or (c) are no longer imputable to lack of knowledge or memory defeats associated with stage (a).

When computerized testing is available, items can be presented either in the usual order (as above), or backwards beginning with stage (c); moreover, specific comments can be delivered according to specific answers, and so on.

Computerized Adaptive and Remedial Testing

Since 1961, the *Service de Mathématiques Appliquées et Traitement de l'Information* (SMATI) at the University of Liège, working under Professor M. Linsman, has been developing Computer-Assisted Instruction (CAI) hardware, software and courseware.

In 1965, DOCEO I, which is discussed more fully by Houziaux (1965 and 1972) and by Lefebvre and Houziaux (1969) was a system equipped with a visual terminal with a screen on which film strips were displayed, view by view, using a 16-mm computer-controlled movie projector. The (multiple choice) answers were introduced through a telephone dial. The computer reacted in a highly-adaptive way through complex conditionings (boolean expressions) either with slides or flashing coded signals. The arrangement is shown in Fig. 4.

In 1972, in co-operation with Professors Van Cauwenberge and Lefebvre from the Faculty of Medicine at the University of Liège, researchers from the SMATI originated

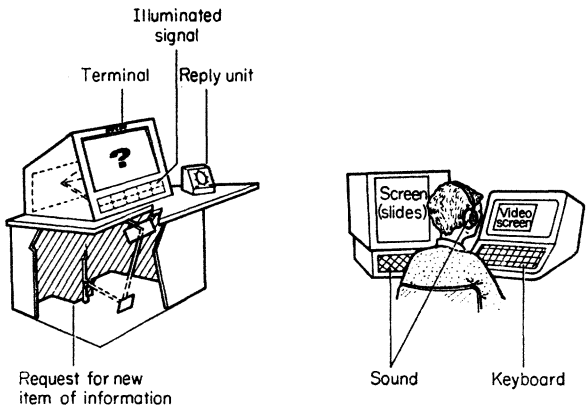


FIG. 4.

the DOCEO II System, which is discussed further in Houziaux *et al.* (1978) and Bartholome and Houziaux (1979). An original special purpose language, known as *Langage de Programmation de Processus Conversationnel* (LPC), was devised for educational and medical applications, the latter involving computer-assisted medical interviewing. This allows, among other things, for the analysis of open answers.

The terminal consists of a cathode ray tube, a keyboard and random-access slide-projector and tape-recorder controlled by the computer. The most important courseware available on this system pertains to medical applications (anamnestic processes for patients with diabetes, and lessons for patients as well as for medical students). Other applications have been produced in a wide variety of fields.

Among the various programmes developed on DOCEO II, I shall describe here only Computerized Adaptive and Remedial Testing (CART) on chemistry, which has been developed by researchers in the fields of chemistry and education at the University of Liège, working under Professor P. Laszlo and Professor G. De Landsheere respectively. This programme presents characteristics from each of the major types of item banks described above. Its originality lies in the combination of various features. The student may, at any time in the programme, ask for further information on any kind of topic within the field tested (here chemistry). He just has to type statements like "What is an isotopic nucleid?" or "I have forgotten the meaning of hologen!" His questions will at once be answered through slides and sound. The (multiple choice) questions of the test, gradually increasing in difficulty, appear in a booklet, one question on each page. After starting with an 'average difficulty' question, each student will progress in accordance with the 'up and down method' or, in other words, with the quality of his answers.

The student is requested to answer as follows. First, the various alternatives appear on the screen as in the following example:

Q. 57	1	2	3	4
%				

The student must fill in the percentage line with the percentage of probability of success he grants each of the solutions. This line must of course total 100, as in the case below:

Q. 57	1	2	3	4
%	20	10	60	10

The up and down rule works in such a way that the next item will be more difficult (progression) if the correct answer has received 50% or more probability of success. But the next item will be easier (regression) if the correct answer has received less than 50% probability of success. Table II indicates by how many items 'up' (more difficult) or 'down' (less difficult) the student will progress or regress according to the probability of success index he has given to the correct answer.

TABLE II

Confidence as to correct answer	0 to 9	10 to 19	20 to 29	30 to 39	40 to 49	50 to 59	60 to 69	70 to 79	80 to 89	90 to 100
Progress in item difficulty	−5	−4	−3	−2	−1	+1	+2	+3	+4	+5
	Down					Up				

Since, in the foregoing example, solution 3 is the correct answer to question 57 and, since the student has given 60% to this solution, he will proceed to a more difficult question, (for example, question 59).

After giving his answer (consisting of a series of percentages), the student is allowed to ask for a piece of information. If he does, he is then asked to answer the same question again, so that the person conducting the experiment can measure how much the student has benefited from the information. Suppose that, after getting the required information, the student attributes 85% probability of success to solution 3 (correct answer); in this case, he will be given question 61. Fig. 5 summarizes part of the progression involved:

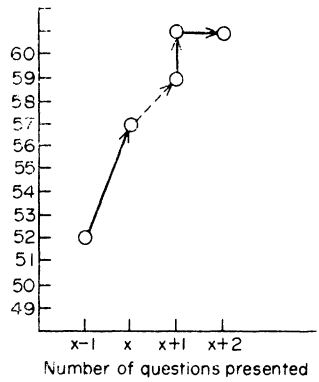


FIG. 5.

The dotted line indicates the progress the student would have made without the help of any information, while the vertical line represents the gain due to the information he has received. This process can be used to assess the gain obtained from comments and feedback in a controlled experiment in which, in contrast to the test group, another set of students receive no comment or feedback whatsoever.

Conclusion

For the last two decades, educational testing has benefited from various improvements. Some are philosophical in nature, such as the will to involve the student in the responsibility for his own learning, a tendency accounting for the various self-regulation

and self-assessment processes. Other improvements are derived from new measurement theories as, for instance, the 'latent trait theory' and item calibration. Last but not least, hardware improvements, and especially the use of computers, open the way to a more individualized and more formative testing approach. Bearing these considerations in mind, it would appear that the time has indeed come to unite these new concepts and tools with a view to increasing their service to the educational cause.

REFERENCES

BARTHOLOME, E. & HOUZIAUX, M. O. (1979) *SIAM-DOCEO II Instruction Manual.*

CHOPPIN, B. (1978) Item Banking and the Monitoring of Achievement, Research in Progress, NFER Series, April.

CRONBACH, L., GLESER, G., NANDA, H. & RAJARAINAM, N. (1972) *The Dependability of Behavioral Measurements: theory of generalizability for scores and profils* (New York, J. Wiley).

DEBOT, F. & LECLERCQ, D. (1978) *Système de Traitement automatique d'Evaluation Pédagogique (STEP)—Guide introductif* (University of Liège).

FISCHER, G. (1975) *Some Probabilistic Models for Measuring Change,* paper given at the 2nd International Symposium on Educational Testing, Montreux.

FISCHER, G. (1977) *Tailored Testing on the basis of the Rasch Model,* paper presented at the 3rd International Symposium on Educational Testing, Leyden.

HOUZIAUX, M. O. (1965) Les fonctions didactiques de DOCEO, in: *Actes du XII Colloque de l'Association internationale de Pédagogie Experimentale de Langue Française* (47–71) (University of Caen).

HOUZIAUX, M. O. (1972) *Vers l'enseignement assisté par ordinateur* (Paris, Presses Universitaires de France).

HOUZIAUX, M. O., GODART, C., LAVIGNE, M., BARTHOLOME, M., LUYCKX, A. & LEFEBVRE, P. (1978) Une expérience d'enseignement assisté par ordinateur chez des patients diabétiques insulinodépendants, *Scientia Paedagogica Experimentalis,* 15, pp. 215–250.

LECLERCQ, D. (1977a) Sequential adaptive tailored testing and confidence marking, in: VANDERKAMP, LANGERAK & DE GRUYTER, *Psychometrics for Educational Debates: Proceedings of the Third International Symposium on Education Testing,* p. 306 (Leyden).

LECLERCQ, D. (1977b) *Concepts, Procedures and Coefficients to be used with Confidence Marking,* paper presented at the 8th European Mathematical Psychology Meeting, Saarbrücken.

LECLERCQ, D. (1977c) *L Matrices or the Computation of Consequences for Confidence Marking Procedures in Educational Settings; Rationale, Algorithm and FORTRAN Program,* paper presented at the 6th Research Conference on Subjective Probability, Utility and Decision-Making, Warsaw.

LECLERCQ, D. (1979) *Test-retest Replication and Spontaneous Acuity of Subjective Probabilities; results from a guessing game,* paper presented at the 7th Research Conference on subjective probability, utility and decision-making, Göteborg.

LECLERCQ, D. (1974) Banques de questions et indice de certitude. Options docimologiques adaptées à l'enseignement secondaire (1re partie), *Education,* No 149, December, pp. 49–59.

LECLERCQ, D. (1975) Banques de questions et indice de certitude. Options docimologiques adaptées à l'enseignement secondaire (2e partie), *Education,* No 150.

LECLERCQ, D. (1976) La fonction régulatrice de l'évaluation vue sous l'angle de l'implication de l'étudiant, *Education,* No 159, December.

LECLERCQ, D. (1978a) Un module d'auto-évaluation ou Comment impliquer l'étudiant dans la régulation de ses apprentissages, *Education*, No 165, February, pp. 59–73.

LECLERCQ, D. (1978b) l'Auto-évaluation des compétences dans le domaine cognitif, *Revue, 13e annee*, No 2, February, pp. 3–20.

LEFEBVRE, P. & HOUZIAUX, M. O. (1969) Anamnèse assistée par ordinateur en diabétologie. Résultats préliminaires, *Revue Médicale de Liège*, 24, pp. 803–809.

LORD, F. (1970) Some Test Theory for Tailored Testing, in: HOLTZMAN W. (Ed.) *Computer-assisted Instruction, Testing and Guidance*, pp. 139–183 (New York, Harper & Row).

LORD, F. & NOVICK, M. (1968) *Statistical Theories of Mental Test Scores* (Addison-Wesley).

NASOETION, N., DJALIL, A., MUSA, I. & SOELISTYO (1976) *The Development of Educational Evaluation Models in Indonesia*, International Institute for Educational Planning (UNESCO).

RASCH, G. (1960) *Probabilitistic Models for some Intelligence and Attainment Tests* (Copenhagen, Danish Paedagogische Institut).

SPADA, H. & LUCHT, H. (1977) *A Situation Test to assess Attitudes: an analysis of the reactions to open-end items based on the model of Rasch*, paper presented at the 3rd International Symposium on Educational Testing, Leyden.

TOURNEUR, Y. (1975) *Effet des objectifs dans l'apprentissage* (Brussels, Organisation des Etudes).

TOURNEUR, Y. & CARDINET, J. (1978) *Des tests à référence normative aux tests de maîtrise par la théorie de la généralisabilité*, SEMME, doc. 780–418, State University of Mons.

WOOD, R. (1975) *Computerized Adaptive Sequential Testing*, dissertation abstract, Chicago University.

WRIGHT, B. & STONE (1979) *Best Test Design* (Chicago).