UNIVERSITE DE LIEGE (Belgium)
Laboratoire de Pédagogie expérimentale

ABSTRACT

L PAY-OFF MATRIXES FOR CONFIDENCE MARKING

or

The computation of consequences for confidence marking procedures in educational settings : the rationale, the algorithm and the FORTRAN program.

D. LECLERCQ (Dr. Educ.)

Abstract of the Paper presented for the Sixth Research Conference on Subjective Probability, Utility and Decision Making, Warszawa, Septembre 1977.

Two waves (in the twenties and in the sixties) of researches on confidence marking have failed to demonstrate the real relevance of the approach. This is due to lacks in the procedure, and specially to the use of unadequate pay-off matrixes.

The instructions given to students must be carefully written, i.e. must present a probabilistic scale of confidence indexes and not only an ordinal one.

The computation of the pay-off matrix has a crucial importance. D matrixes (matrixes computed according to decision theory) must be used. Moreover, other properties are helpful for everyday education use.

It seems convenient to define A, E, D, I, and L matrixes, where only D, I and L are acceptable, and where L ones are strongly recommanded.

Superiority of the D matrixes can be shown from mathematical and graphical evidences. A computer program can compute by interative steps the desired L matrixes for various situations. Typical examples are given.

In conclusion, it is claimed that, provided the experiments are based on methodolically admissible procedures, confidence marking can provide new relevant information. A third generation of researches on confidence marking should now start.

leiden
the
netherlands
june 27-30 1977

third
international
symposium
on
educational
testing

SEQUENTIAL ADAPTIVE TAILORED TESTING
AND CONFIDENCE MARKING

by Dieudonné A. LECLERCQ, Ed. D.

Chercheur au Laboratoire
de Pédagogie expérimentale
(prof. G. De Landsheere),
Université de Liège,
Belgium.

Chargé d'enseignement
aux Facultés N.D.P.
de Namur, Belgium.

In order to present to a student the most adapted question (or sequence of questions) through a computer terminal, a familiar way consists in doing two things. First, one has to create an item pool with an appropriate structure. Second, one has to develop an algorithm (or a list of sequential rules) that will select an item from the pool depending on the student's previous reactions.

Organisation of the item pool

Some anthors (F. LORD, 1970, G. FISCHER 1975) recommand to rank the questions on a RASCH-like scale (RASCH 1960) so that the difficulty level of an item can be expressed by a "population free" index. In such a model, each item has its'own (item caracteristic curve) (ICC) and classical indexes are replaced by others, the distribution curves being logistic ones, from the general equation.

$$P_i(\theta) = C_i + \frac{1 - Ci}{1 + \exp\left[-1.7\, ai\, (\theta - bi)\right]}$$

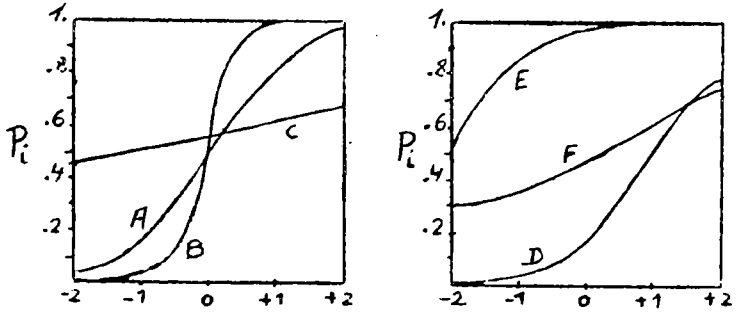
for $-\infty < \theta < \infty$
(θ = student ability)
from LORD 1970, p.142

- pi (difficulty estimated by frequence of good response) is replaced by bi (difficulty corresponding to the inflexion point of the ogive-like curve) varying from $-\infty$ to $+\infty$ (1).

- di (discrimination estimated by point biserial correlation for instance) varying from -1 to +1, is replaced by a_i (discrimination corresponding to the slope of the ICC) varying from 0 to ∞ (2).

- gi (chances of success due to random guessing) varying from 0 to 1, is replaced by C_i (determined by the lower asymptote of the ICC) varying from 0 to 1. This value is quite useful for multiple choice questions.

This three parameter logistic function gives the following kind of curves.



(1) Transformations on this axis has few importance, so we can consider that, for practical use, $-3 < bi < +3$ (F.LORD 1970, p.147).

(2) "We plan to use items that are positively correlated with θ" (LORD, 1970).

Values :	A	B	C	D	E	F
a_i	0	0	0	1	-2	1
b_i	1	3	0,2	1	1	0,5
c_i	0	0	0,33	0	0	0,25

Recently B. CHOPPIN (1975) has proposed to express b in wits on a scale varying from 0 to 100.

Building an appropriate algorithm

According to a common reasonning (1), one shall try to present at time v to the student a question so that he will succeed with Probability (P = 50). Commonly, in the up-and-down method (LORD 1970, WOOD, 1975), the choice of the item will in fact be restricted to the choice of the best bi at time v. Assuming the Markov property (2), this choice depends only of bi at time v = 1, being a constant.

A first problem arises with determining θ, how hypothetical or temptative it may be.

When an item has already been answered, likelihood function could be used. But, as LORD notes, "this complicates the problem to such a point that we shall not attempt to evaluate the results obtained when the stockastic process itself depends on successive maximum likelihood estimation" (p.148).

(1) On which DAVIS' rule of item selection was based (reject them with pi $<$ 27 % and pi $>$ 72 %.

(2) After b(v) is given the probability of any value of b(v) is independant of values anteriors to b (v-1).

The result of such difficulty is that actually, we starts with an item that has bi = 0 (the center of the scaled items). It looks like of Laplace's strategy in decision theory : equiprobability or complete ignorance. We shall try to present an answer to this first problem with the student's general estimation.

The second problem concerns step sizes, that are, usually, determined in advance. Ideally, it should change according to the student's response caracteristics (accuracy , rapidity, confidence...). A mode of responding that permits individualised step sizes during the interaction of the student and the item pool, is proposed hereafter.

Description of the experiment

In a multimedia project (Prof. DE LANDSHEERE, LASZLO, Univ. of Liège, Belgium), 113 multiple choice item have been ranked in a booklet (1) from page 1 to 113. Four such booklets were available (2).

Each student is invited to start with item 50 (almost the center of the booklet). A grid appears on the screen as follow :

Question 50

Alternatives	1	2	3	4	5
Chances in %					

(1) These question on a college chemistry course have been prepared by DONNAY, GREANJEAN, VANBELINGEN, CORNELIS, STOCKIS and LECLERCQ.

(2) The questions were ranked according to a branching structure of behavioral objectives and to the frequence of success on previous trials.

The student using a keyboard has to distribute his 100 % chances on the various alternatives plus one, not on the screen (1).

For example, here is an acceptable pattern of response :

0 10 50 0 0 40

because it sums up to 100, and the number of answers is equal to n or n+1 (here 5 or 6).

The computer (2) was programmed so that patterns could be accepted for summing up going from 95 to 105.

Those procedures have been used previously (J.D. BAKER,1968, W.EDWARDS,1967), in experimental settings. In order to use these principles in educational situations, six problems must be specially studied.

1. Student's willingness

If one presents this mode of responding as a common one, students are likely to find this as normal as any other mode of responding.

Of course, it takes a little more time than just choosing the answer and it constitutes a constrain. If they had the choice (but they had not), the students would prefer a simpler method.

---

(1) The students had been informed of the presence of questions without any correct alternative ; in such case, one has to choose n+1, i.e. the number of alternatives plus one. In order to keep the students vigilant and discourage guessing, this possibility is not reproduced on the screen.

(2) This program has been implemented on a console called DOCEO II, developed by HOUZIAUX, BARTHOLOME (S.M.A.T.I., Univ. of Liège) in L.P.D. (oriented) language. The audiovisual components of this terminal will be explained in an other paper. The host computer is a PDP/8.

2. Students estimating ability

This is the crucial point. The experiment has been conducted after theoretical and experimental research on confidence marking procedures and students behaviors in such situations (LECLERCQ, 1975).

It appears that adults can,in familiar situation, discriminat up to nine or ten (1) steps of confidence on the scale. For this reason, only ten points were considered on the continuum(each portion including 10 % chances), in the computer program.

Research had shown (DE FINETTI ,1965) that adults can do this validly (2) only after specific training with immediat feedback (3).

That is why the scores (and not only the good response) were displayed immediately after response.

3. The scoring rule

Some of our previous research it appeared that not the use of pay off matrix is not advisable when dealing with confidence marking indexes. The must be concieved in such a way as to arrange "admissible probabilities measurement procedures (see SHUFORD, ALBERT and MASSENGILL (1966)).

---

(1) One more "magical number seven" for MILLER (1956)'s collection.

(2) Avoiding the unwanted strategies described by decision theory.

(3) Communicating the real consequences immediately and repeating the process.

The matrix was computed on equation and assamptions that are frequently met in subjectivists' works (SHUFORD, RAIFFA, VAN NAERSSEN, DE FINETTI, etc.).

With the following matrix, the best strategy is saying the truth : it maximises the expected value of the item score and, of course, of the total test score.

0 - 9 %	10 - 19 %	20 - 29 %	30 - 39 %	40 - 49 %	50 - 59 %	60 - 69 %	70 - 79 %	80 - 89 %	90- 100%
0	9	13	20	23	24	26	29	30	31
0	-1	-2	-5	-7	-8	-11	-18	-22	-31

This is a D matrix (1) and not a Q one (2). More over, it has only the lowest possible integers : it is an L matrix.

Using a D matrix appeared to be one of four methodological rules that must be followed to measure validly the students' confidence.

Let us compute a student's score with the following pattern :

Alternatives	1	2 *	3	4	
Chances	0	60	30	10	
Score	0	26	-5	-1	total = 20 points.

(1) In accordance with Decision theory. It is not possible to explain the matrix computation here.

(2) In french : une matrice quelconque = any matrix.

4. The branching rule

Fundamentally, we have used the"up-and-down method", but the size of the steps depended both on the objective quality of response and on the subjective estimation. Two different rules were used :

Rule 1

When the chances given to the good alternative are greater than 50, step sizes are as follow :

Chances	Step size
51 - 60 %	+ 1
61 - 70 %	+ 2
71 - 80 %	+ 3
81 - 90 %	+ 4
91 -100 %	+ 5

When the chances given to the good alternative are lower or equal to 50, decreases are as follow :

Chances	Step size
41 - 50 %	- 1
31 - 40 %	- 2
21 - 30 %	- 3
11 - 20 %	- 4
0 - 10 %	- 5

Rule 2

Fifty percents given to the good alternative remains the critical point for decrease or increase in item difficulty, but an exponential-like rule is used :

Chances	a	Step size
51 - 60 %	+ 0	$2^a = 1$
61 - 70 %	+ 1	$2^a = 2$
71 - 80 %	+ 2	$2^a = 4$
81 - 90 %	+ 3	$2^a = 8$
91 -100 %	+ 4	$2^a = 16$
On the other side,		
41 - 50 %	0	$-(2^a) = -1$
31 - 40 %	- 1	$-(2^a) = -2$
21 - 30 %	- 2	$-(2^a) = -4$
11 - 20 %	- 3	$-(2^a) = -8$
0 - 10 %	- 4	$-(2^a) = -16$

5. The starting rule

Unfortunately, we have not experienced the idea of requesting the student to estimate his level at the total test, so that we could have placed him at this most likely level.

6. The stopping rules

The were a lot of such rules in this experiment. If students got bored, they were allowed to leave two out of sixty did so before the end). After 10 questions or one hour, each student was stopped. A few students made sobig steps that they reached the ailing (question 113) before the end.

Results

A great variety of types of progression were observed. They were complicated by audio-visual informations given on the student's request.

Since a pattern of responses was given by the student before and after each information, it was easy to measure the effect of each message. Not only the changes of probabilities given to the right alternative have been recorded, but also where (on wich distractors) changes occured. Results have been interpreted with regard to students'comments on each A.V. message.

In order to obtain distributions of probabilities for each question, other students (40 per item) answered the same way to the 113 questions, but on a paper-pencil mode.

Conclusion

Rised tailored testing resquests accurate rules(scoring, starting, branching and ending rules), based on sophisticated responses for it appears that the students are abble to give them and we are abble to measure them. The first exploration of the possibilities of confidence marking is encouraging and indicates that research should go on in this direction.

D. LECLERCQ.