Psychology Teaching Review (1995) 4, 2

Human experts and computerized procedures in knowledge assessment

Maria Kambouri, Institute of Education, EPSEN, University of London and **Karl R. Gegenfurtner**, Max-Planck-Institute for Biological Cybernetics, Germany

The assessment performance of human experts was compared with that of a simple, uni-dimensional, computerized procedure on a standardized task. Experienced tutors were asked to predict the responses of students (correct or incorrect) to some items of a standard high-school-mathematics test, on the basis of some previously observed responses to other items. Both the human experts and the computer procedure selected items, observed the responses, and made predictions concerning the remaining items. This sequence was repeated several times. Detailed comparisons showed that the computer algorithm performed at least as well as the experts both in selecting the most informative items and in predicting student responses.

Introduction

In recent years the use of computers in the classroom to assist with instruction has become widespread. However, the assessment of a student's knowledge is mostly carried out by teachers via oral examinations or standardized written tests. Although a face to face examination is still considered the best way of assessment in most fields, it has at least two important limitations: it can be very time consuming for the teacher in the average size classroom and it may prove too stressful for certain students. Standardized tests on the other hand not only are insensitive to individual differences but provide no immediate feedback which is so important for a positive assessment. Computerized assessment may well be the solution. Computer algorithms for knowledge assessment can be tailored to suit individual students' needs and can reduce the amount of time needed for testing while being a less intimidating examiner.

Providing immediate and appropriate feedback to each student is a primary advantage of computerized assessment and instruction. The benefits of responsive feedback have been recognized in early attempts at computer-assisted instruction (see for example, Suppes, 1966). In fact, issues concerning the design of appropriate feedback date back to psychology's earliest efforts to examine the process of learning (Thorndike, 1913).

Indeed, the introduction of computers in education and the rapid evolution of information technology has begun to influence tests and ways of testing. Alternatives to standard testing are being used in classrooms and are reported to have met their goals (see for example Bork, 1987; Fetta and Harvey, 1990). However there is little research on how interactive assessment by computer algorithms compares to human expert assessment.

In this paper we propose to illuminate this issue through a detailed analysis of the comparative assessment performances of human examiners and computer algorithms. In particular, we are interested in questions of the following type: how much information

has been acquired by the examiner at each step of the assessment procedure? Is this information about the knowledge of the student used efficiently to select the most appropriate item on which to examine the student? How well can an experienced teacher judge the future performance of a student based on responses to a few questions? And, finally, can assessment be carried out as effectively by computer algorithms?

To investigate how close computerized assessment is to teacher assessment we compared the performance of a computerized procedure to the way highly experienced teachers (experts in their field) performed on a standardized assessment task.

Methods

In a typical educational setting the teacher would ask the student a question and would then adjust his or her opinion about the student's knowledge according to the student's answer. This interaction served as a guideline for our experimental task. The teacher was asked to select appropriate test items in mathematics to ask the student based on the information received about the student's performance on previously asked items. Next, the teacher was required to predict whether the student will solve subsequent items.

Six highly qualified teachers with an extended experience to interact with students in one to-one settings served as the experts in our experiment. The students' data came from the Regents Competency Test (RCT) in Mathematics, a New York State examination taken in June 1987 in New York City. In particular, we used the 20 open-ended problems (coded as correct or incorrect) from a random sample of 16 out of the 7,387 students who failed the test. All the experts were highly familiar with the RCT test and were informed about the student population.

Each expert was presented with a copy of the 20 RCT problems and was instructed to assess one student at a time on an interactive computer session. At each step of the procedure, the expert was required: (i) to select 3 problems, one at a time, to "ask" the student and, having received the student's answers to these 3 items (ii) to predict the student's performance on the non-selected problems which remained. For each item the experts were required to predict whether or not the student would solve the item, and then give a confidence rating about how certain they were. This was repeated three times. In all, there were 3 sets of selections and 3 sets of predictions for each student. Performance was therefore measured on 17, 14 and 11 of the 20 test items. Experts spent, on average, 25 minutes to assess one student through this computerized task and, at the end of each session, they received feedback on their perfomance.

Results

Predictions

As a measure of assessment accuracy we recorded the number of correctly predicted items for each expert at the end of each step. The results are pictured in Figure 1. The bars show, for each expert, the proportion of correct predictions in all three steps of the task. The column on the far right shows average performance. For all experts, performance was significantly above the 0.5 chance level ($p<0.05$). However, results indicate that experts perform rather moderately in the assessment task: performance varied between 59 per cent to 72 per cent correct predictions (at the end of each session). To test whether performance increased significantly from one step to the next, we used a likelihood ratio test which accounts for order restrictions (see Barlow et al., 1972, p.119). We tested the null hypothesis of equal mean proportions in each step against the alternative of ordered means: $\mu_1 < \mu_2 < \mu_3$, where μ_i is the mean proportion in step i. Even though only two experts (B and C) showed a noticeable improvement over steps, the average performance increased significantly. We therefore conclude that, in general, experts do make use of the accumulated information about the student's performance (correct/incorrect response), which they received by "asking" questions (selecting items to see the answer to) at each step of the assessment task.

A detailed analysis of performance by item and by student showed that there was no systematic relationship between either item difficulty or student ability and performance by the experts. There was, however, a significant correlation between the experts' predictions: with only one exception

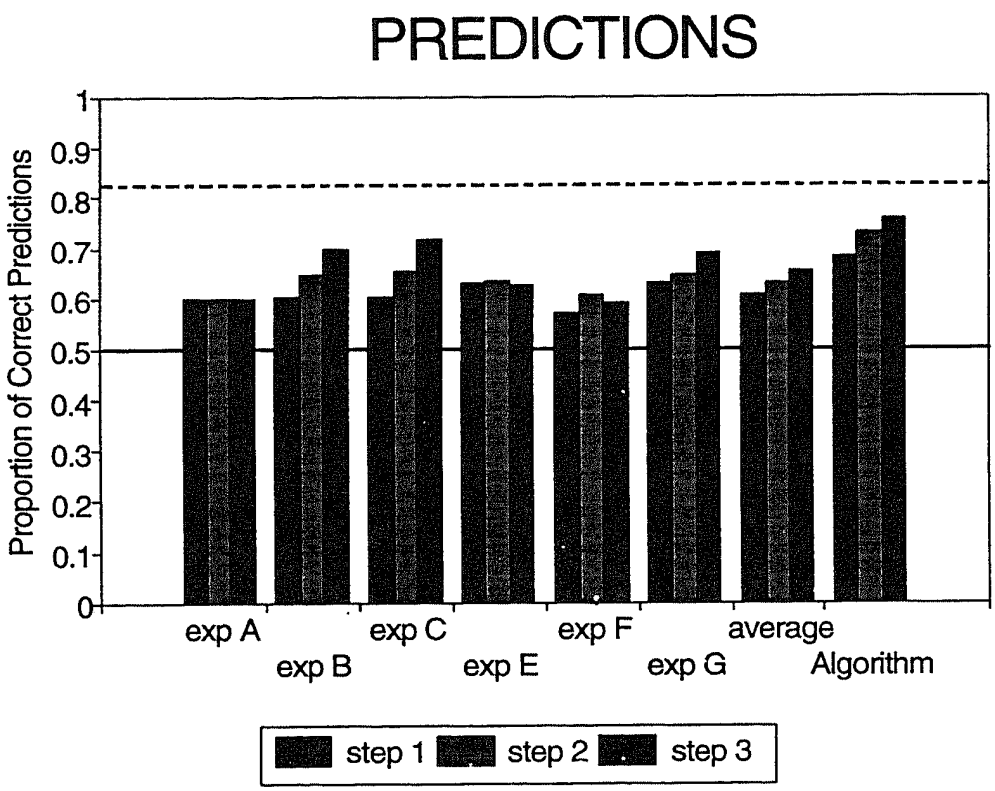

Figure 1. Proportion of correct predictions for each expert, the average over all experts, and the linear computer algorithm. Performance is shown at the end of steps 1, 2 and 3

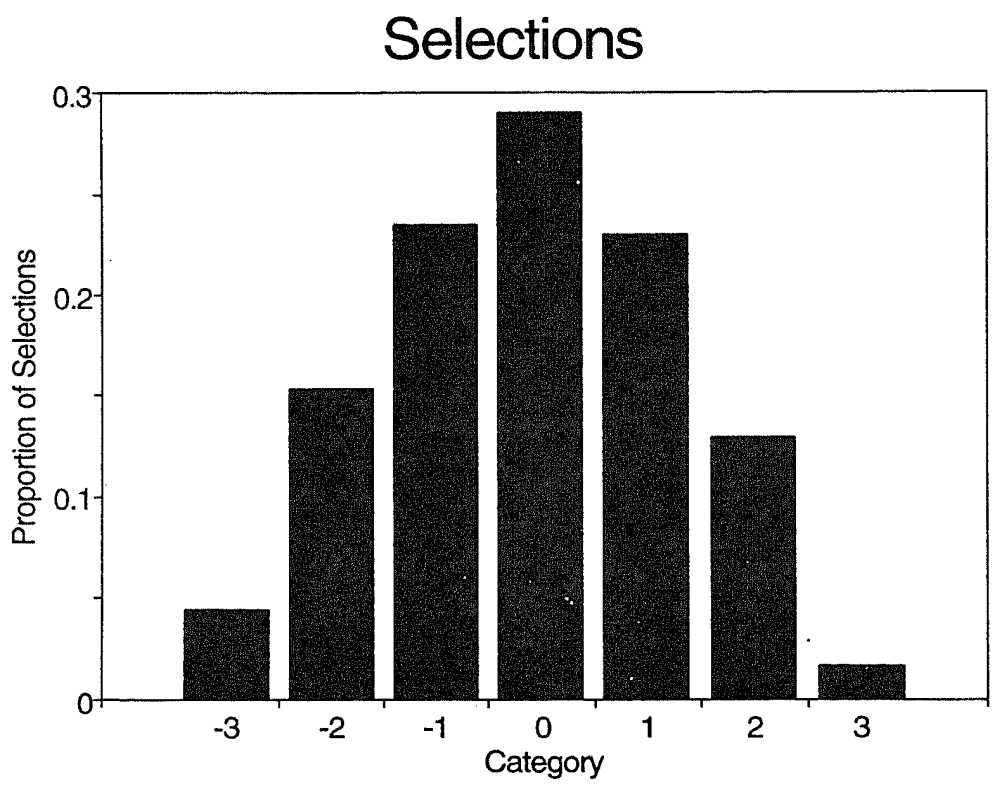

Figure 2. Category usage averaged over all experts. The number of items that were selected from each category was divided by the absolute number of items in that category. Experts mostly selected items about which they were uncertain

(experts G and F) all correlations calculated on each pair of experts' predictions were significant and positive. This result shows that experts shared inefficient prediction strategies and tended to make similar errors.

Selections

Expert assessment performance on the task was analysed further through the selection processes. One way to investigate whether experts' selection strategies were efficient is through the confidence ratings they gave about their predictions. If experts follow an efficient strategy, they would be expected to "ask" those items which they were least sure about when predicting the student's performance. On the other hand, experts may be using an inefficient strategy of asking those items for which they were fairly sure – in order to verify their intuition about the student's knowledge. Therefore, we looked at how experts had rated those items which they subsequently selected to ask the student. For each confidence category we calculated the frequency with which an item in that category was selected at a later step of the procedure. To account for differences in category usage the number of selected items in each rating category was divided by the total number of items in that category. Because after the above normalization, experts have quite similar item selection patterns over the different categories, we aggregated data over all experts (Figure 2). In general, experts mostly selected the items for which they had to guess (category "O") the student's performance. If experts had followed a random selection strategy, all bars would have been of the same height. Therefore experts did adopt an efficient strategy and mostly selected items from category "O".

Algorithm

Having seen how experts perform in this assessment task, we wanted to test whether a computer algorithm could do as well as the experts. Our goal was not necessarily to find the "best" algorithm but to find a simple algorithm that would stand a comparison with the experts. It is common in psychological research to use some representation of the concept of "ability" in order to analyse the results of mental tests (Lord, 1952; Lord and Novick, 1968; Weiss, 1983). The linear algorithm which was used here was indeed motivated by unidimensional Item Response Theory (IRT) and tailored testing methods (e.g. Guttman, 1944; Rasch, 1960; Thissen and Steinberg, 1986), in the sense that student ability and item difficulty were represented on a single scale.

The algorithm was designed to derive its predictions about the students' future performance directly from student data obtained on the RCT test. It was programmed to carry out the same assessment task given to the six experts. The following is an outline of the algorithm: According to a *questioning rule* an item is selected to "ask" the student. The information on the recorded answer is then used through an *update rule* to improve the estimate of the student's ability. After the three items in the first set have been selected, the algorithm's *prediction rule* predicts the student's performance on the non selected items.

Our assumption that student ability and item difficulty can be represented on a common scale allows the simple prediction rule that a student will solve an item if his or her ability is greater than the item's difficulty. We assume that the algorithm has information about the difficulty of each item, and about the average ability of the student population. The algorithm can then estimate a particular student's ability from the available data in the following way. Let us denote the student's answer x_{si} by 1 if student s solves item i, or 0 if the student does not solve item i. Let N be the number of students and M the number of items in study. We then define item i's difficulty, d_i, as follows:

$$d_i = 1 - \frac{1}{N} \sum_{s=1}^{N} x_{si}$$

d_i provides an estimate of the proportion of students failing item i and is calibrated on a large sample of 7,387 student answersheets taken from the population of reference. At trial n = 0, we start by letting the student's ability be equal to the average student ability in the population, that is,

$$a^{(0)} = \frac{1}{NM} \sum_{s=1}^{N} \sum_{i=1}^{M} x_{si}$$

In this sample of 7,387 student tests the mean score was 0.58, thus $a^{(0)} = 0.58$. Then, at

any trial n, (n = 1, ...) the three basic rules necessary to carry out the assessment task for the student under examination are the following:

Questioning-rule
Select item i (i not in P_n) such that $|di - a(n-1)|$ is minimized. This way the algorithm selects the item which is hardest to predict (since it is closest to the student's ability). Define $P_{n+1} = P_n \cup \{i\}$, where P_n, is the set of all items selected up to and including trial n, (thus P_0 is the empty set). We write $|P_n^+|$ for the number of items in P_n solved by the student and $|P_n|$ the total number of items in P_n.

Update-rule
For some fixed parameter β, $0 \leq < 1$, compute

$$a^{(n)} = (\beta)\, a^{(0)} + (1-\beta)\frac{|P_n^+|}{|P_n|}$$

If β has a value of 1 the predictions will be the same for all students, based only on the average student ability $a^{(0)}$. If β is 0, then the predictions will be based solely on the student's responses so far. Whereas we originally started out with the later strategy, empirical testing showed that using a weight $\beta = 2/3$ (a strong bias towards the mean score) was an advantage. This value led to better results than other weights (e.g. "0", "1/3", "1/2", "1") which were tried out. The main advantage of this correction is that it avoids large changes in the estimate of $a^{(n)}$ during the first few trials.

Prediction-rule
Predict item i to be correct if $a^{(n)} \geq d_i$, or incorrect if $a^{(n)} < d_i$.

With these three rules the algorithm was now performing the same task as the expert and therefore a direct comparison was possible. The results of the application of these rules for the same data presented to the experts are discussed below in terms of predictions and selections.

Predictions
The proportion of correct predictions at each step of the procedure made by the linear algorithm is shown by the bar on the right hand side of Figure 1. The predictions made by the linear algorithm are systematically better than those made by experts. Using a z-score test for differences in proportions we rejected the null hypothesis that performances for the algorithm and the experts were equal at the 0.05 significance level. The algorithm's performance was compared to each of the experts individually. At each step of the task the proportion of correct predictions made by each of the experts was consistently lower than the algorithms.

Predictions by experts show, on average, improvement over steps. A similar result was obtained when the algorithm was implemented on a large sample of student data. A chi-square test indicated a highly significant improvement over steps (x^2 = 1274.8 with 2 degrees of freedom).

Selections
Given the algorithm's prediction and selection rules described above the range of items to select is much narrower than the experts'. This algorithm always begins assessment with the same item (the one closest to the average ability), and because of the weight 6 it selects mostly items with a difficulty higher than $(2 \times a^{(0)}) / 3 = 0.38$ and lower than $(2 \times a^{(0)} + 1) / 3 = 0.72$. This selection rule avoids items which a large majority of students solved or failed, and which are therefore easier to predict.

Because of these differences in selection strategies, it is unclear whether the model's superior performance is caused by better selections or better predictions. Separating the two modules in a way which permits a comparison of performances on predictions only can be achieved by keeping selections unchanged. For this, the linear algorithm was re-implemented using the problems selected by the human assessors. Thus both assessors have at each trial the same information available (the same set of items for each student examined) and the comparison involves the way each of them uses this information to make predictions. The linear algorithm, modified in this fashion, was still significantly better than the experts, but its performance was somewhat lower than before (X^2 = 26.03 with 2 degrees of freedom). Results of this comparison support the previous finding that the model's predictive power is superior to the experts' in this task, Furthermore, a better selection strategy alone cannot explain the differences in per-

formance. The linear algorithm's superior performance derives from the selection as well as the prediction module.

The performance achieved by the linear algorithm is significantly above the experts, but it is still rather low on an absolute scale. However, since our goal was merely to report the existence of such a simple algorithm, we did not search for further improvements. The choice of this particular model was dictated by programming simplicity, predictive power and algorithmic efficiency. We did however compute an upper bound for one-dimensional algorithms that code ability and difficulty on a single scale. Performance of this "optimal" algorithm is indicated by the dashed line in Figure 1 and was at 82 per cent correct predictions. Thus, performance of the linear algorithm and of the experts is quite close to being as good as possible assuming only one underlying dimension. It is unclear whether the remaining 18 per cent of performance could be achieved using multidimensional algorithms, or whether it is simply due to noisy data, caused for example by students making careless errors or lucky guesses.

Discussion

In contrast to paper-and-pencil tests, examination in one-on-one settings enables the teacher to effectively reduce the number of questions to ask the student, by making inferences from the student's answers to previous questions and by selecting more appropriate questions to ask. This form of assessment leads to a shorter and more in depth testing of the student's mastery of that material and provides immediate feedback which is vital to the learning process. As in most of today's schools it would seem unrealistic to expect such thorough testing to be carried out in the classroom, the solution of adopting computerized assessment to assist teachers in their tasks comes naturally. But how efficient and accurate is a computer algorithm in assessing a student's knowledge of a given topic? How does it compare with the teacher-student examination? Our results allow us to answer these questions. Assessment of a student's mastery of a selected topic was investigated in a controlled assessment task carried out by expert teachers of maths. We compared the expert's performance on this task to the performance of a linear algorithm programmed to perform the task in a similar way.

One significant component in the assessment procedure is the updating of inferences about the student's mastery of the topic made by the teacher every time he or she asks a new question. The first issue we examined therefore was how information provided to assessors at each step of the assessment procedure was handled. In other words, whether there was any evidence that experts incorporated the accumulated information about the student's knowledge in their assessment. By considering the proportion of correct predictions made at each step, we found that expert performance between steps increased significantly. We thus showed that experts' judgement of the student's mastery of the topic became more accurate as they examined the additional information at each step of the procedure. The computer algorithm's performance also increased from one step of the assessment procedure to the next.

Another pivotal component in the assessment process is the careful selection of the right questions to ask so as to minimize time and converge to an accurate state of the student's knowledge. We therefore explored the efficiency of experts' item selection strategies: this was achieved through examination of the ratings which accompanied their predictions. Experts most frequently asked those items about which they were least sure when predicting the student's performance, thus showing an efficient selection process. However, they were not quite as consistent in adopting this strategy as the linear algorithm was, whose selection rule was designed to always select the most uncertain item.

Overall, these results show that the algorithm's performance was in most cases significantly more efficient than the experts', both due to a higher percentage of correct predictions and a more consistent strategy of selecting the most informative items. The superiority of the linear algorithm should not come as a surprise as it is in line with research on human judgment carried out by several investigators. For instance, Goldberg (1968) compared undergraduate students and experienced clinical psychologists and psychiatrists in their ability to diagnose psy-

chosis using the Minnesota Multiphasic Personality Inventory (MMPI). He found that experts performed at 65 per cent, only slightly better than novices-students (58 per cent) or random (50 per cent). When a simple linear regression model was used to predict the same data it made accurate diagnoses 70 per cent of the time. Other cases where such models have outperformed human judges are reported, for example, a model of predicting success in graduate school (Dawes, 1971). Such results have led to the comparison of experts to even simpler models which use equal weights for the important variables. These models, which have been called *improper linear models* because they do not involve statistical estimation, are often superior to experts (Dawes, 1979). Indeed, in that study, a simple one-variable model predicted business failure more accurately than 31 of the 43 experts.

The problem of human judges not being able to incorporate background information in their judgement is known in the cognitive literature as the distinction between base-rate versus case-specific data (Kahneman and Tversky, 1973). In a series of studies, it is argued that "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations". According to these studies, the inability to use base-rate information is therefore the key to the relative weakness of experts, as compared to models. In a study of decisions made under uncertainty, Johnson (1988) pointed at a rather pessimistic appraisal of experts in making correct diagnoses. Although he acknowledged experts' strengths at the interpretation of cues that apply to particular cases, he stressed their weaknesses (and the strength of models) in the ability to combine more "mundane" information available for every case. This conclusion is emphasized by our experimentation with the linear algorithm. The parameters that do not vary from student to student, namely item difficulty and average student ability, are base-rate information. As seen above, the linear algorithm still performs better than the experts even when its update-rule (the parameter 6) is adjusted so that only base-rate information is taken into account and the information about the particular student is ignored.

Although the picture painted by empirical research in behavioural decision theory is grim, cognitive science literature which investigates expertise has often taken the other side. In the area of problem solving it is found that human experts have a rich repertory of strategies (and appropriate mechanisms for assessing and applying these strategies) which often allow experts to perform tasks more effectively than novices (for example see Larkin et al., 1980). It must therefore be stressed that the nature of the task assigned to experts is a key factor in this debate.

The main result shown in this work, that computerized assessment can be at least as efficient, and indeed more accurate and consistent than when performed by a human examiner on this standard examination task, is useful for educational purposes. In higher education, where computers already have their place in supporting the teaching and learning, it would be natural to apply computerized assessment as an extension of tailored-teaching for each individual student. In teaching statistical methods for conducting research in psychology, for instance, computers are widely used for analysing data with statistical packages like SPSS. When it comes to examination of this material most colleges adopt either a 3-hour written examination or the submission of a report/essay by which students are required to show mastery of one or two statistical methods. Neither of the two ways of examination are as complete and efficient as a system of ongoing assessment which allows lecturers to monitor student progress throughout the year. This "early diagnosis" of difficulties, coupled with the possibility of immediate feedback, are the best prescription for effective learning.

Acknowledgements

The research described in this paper was part of the PhD dissertation of Maria Kambouri at New York University. The authors would like to thank Jean-Claude Falmagne and Geoff Iverson for valuable comments on an earlier version of the manuscript.

References

Barlow, R.E., Bartholomew D.J., Bremner J.M., and Brunk H.D. (1972) *Statistical Inference under Order Restrictions. The theory and Applications of Isotonic Regression.* New York: Wiley

Bork, A. (1987) *Learning with Personal Computers.* New York: Harper and Row

Dawes, R.M. (1971) A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180-188

Dawes, R.M. (1979) The robust beauty of improper linear models in decision making. *American Psychologist, 34, 7,* 571-582

Fetta, I., and Harvey, J. (1990) Technology is changing tests and testing. *UME TRENDS: News and Reports on Undergraduate Mathematics Education , 1,* 6

Goldberg, L.R. (1968) Simple or simple processes? Some research on clinical judgments. *American Psychologist, 23,* 483-496

Guttman, L. (1944) A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150

Johnson, E.J. (1988) Expertise and decision under uncertainty: Performance and process. In M.T. Chi, R. Glaser and M. Farr (Eds) *The Nature of Expertise.* Hillsdale, NJ: Laurence Erlbaum

Kahneman, D., and Tversky A. (1973) On the psychology of prediction. *Psychological Review, 80,* 4, 237-251

Larkin, J., McDermott J., Simon D.R., and Simon H.A. (1980) Expert and novice performance in solving physics problems. *Science, 208,* 1335-1342

Lord, F.M. (1952) A theory of tests scores. *Psychometrica, Monograph No.7,* 17, 4, Pt.2

Lord, F.M., and Novick, M.R. (1968) *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Denmark Peadagigiske Institut

Suppes, P. (1966) The use of computers in education. *Scientific American, 215,* 206-22 1

Thissen, D., and Steinberg, L. (1986) A taxonomy of item response models. *Psychometrica, 51,* 4, 567-577

Thorndike, E.L. (1913) *Educational Psychology.* New York: Teachers College

Weiss D.G. (Ed.) (1983) *New Horizons in Testing: Latent Trait Theory and Computerized Testing.* New York: Academic Press

Address for correspondence: Dr Maria Kambouri, Institute of Education, EPSEN, 25 Woburn Square, London WC1H 0AA. Telephone 0171 612-6301, email m.kambouri@mentor.ioe.ac.uk.