C h a p t e r Seven

A COMPUTER-BASED TEST ITEM BANK USING A RASCH MODEL*

INTRODUCTION

This chapter describes a computer-based test item bank which I am developing for use in the English Language Centre of the University of Poznań. The system depends on a statistical analysis using a Rasch model, or Latent Trait theory. This is a recent and still relatively little-known approach to statistical analysis, and this paper will thus include an introduction to the probabilistic principles upon which it is based, and a brief discussion of the advantages claimed for it. Various computer programmes have been described for fitting the Rasch model to data. The item bank described uses a computerised version of a procedure for hand-working called PROX. Readers interested in actually using the statistics would need more guidance than the limited space here allows (e.g. Henning 1987). Therefore only the background to the statistics is described in this chapter.

The paper also describes how within the constraints of a machine-based system I have attempted to preserve the possibility of constructing valid test items, and thus avoid a common criticism of computer-based testing.

A BRIEF OUTLINE OF THE SYSTEM

Figure 1 illustrates the system in outline. The central box is the domain of the computer system, while the items around it are in the world outside the computer.

---

* Neil Jones, Adam Mickiewicz University, Poznań.

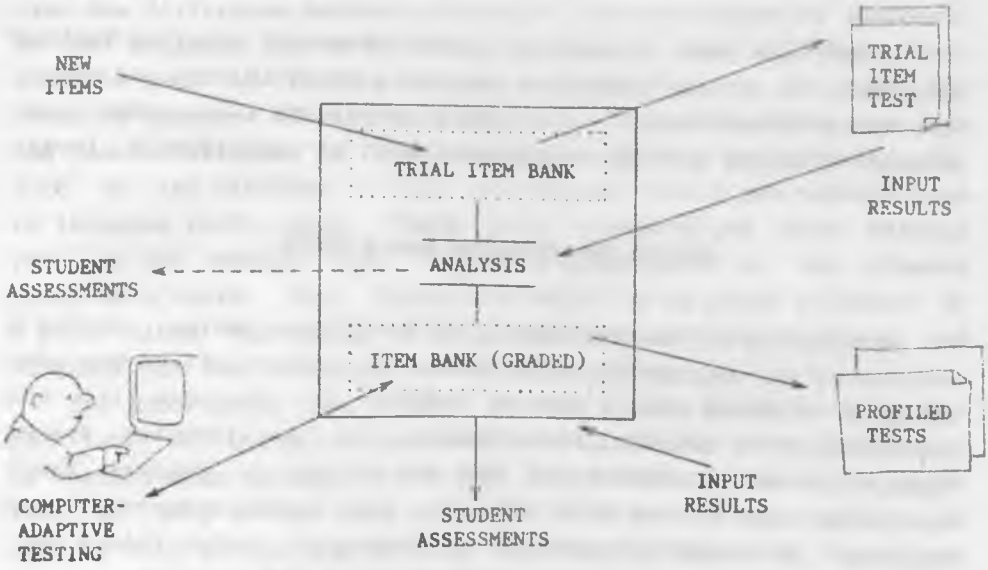

Fig. 1. The system in outline

1. Test items are written onto the computer and saved in a small bank of trial items.

2. A trial test is printed using some or all of these items. If there are items in the main bank some of these will be used too.

3. This trial test is given to a sample (at least 100) of students, marked, and the results input.

4. The results are analysed. Estimates of candidate skill level and item difficulty level are produced, together with estimates of goodness of fit. Items and candidates who fit the model badly, i.e. perform erratically, are identified and bad items are rejected. Good items are passed to the main item bank where they are stored in rank order of difficulty.

5. This process is repeated and the item bank grows. All items are located on the same scale of difficulty.

6. The bank is used for generating and printing tests with specified profiles: difficulty level and range, language area. These tests can be used for specific purposes: placement, achievement, diagnosis.

7. The same bank is used by a CAT (Computer Adaptive Testing) program in which the candidate interacts directly with the computer, which selects items according to candidate's previous responses, to arrive quickly at an assessment of candidate's level.

REASONS FOR DEVELOPING SUCH A SYSTEM

My interest in testing begins with syllabus design, within a methodological approach broadly known as 'Communicative'. The name reflects a belief that a central concern of language is the transmission of meaning, that learning of the language system tends to occur in students as they are helped to express their own ideas, that fluency will naturally come before accuracy, that learning is a creative process of discovery, rather than a passive process of obedient submission to drill and practice.

It is difficult to write a syllabus for a Communicative language course. The traditional structural syllabus -- a list of grammar and vocabulary items, ordered -- has the disadvantage that it imposes on the learners a progression that may well not match their internal learning map, and may thus be in large part irrelevant. More recent attempts at syllabus design, where units of planning are language functions, or semantic notions, or operations that students will be able to perform in the target language, tend in practice to be realised as structural programmes, heavily disguised and disordered. I tend to believe that a truly communicative syllabus should be essentially a description of pedagogically useful, interesting classroom activities (i.e., at least as much a *methodological* statement as a language content statement) planned with some consideration for what is linguistically practical at each level. Engaging in these activities, each student learns in his own way and at his own pace, but efficiently.

The notion of 'linguistically practical' recognises that activities must be graded to suit students' level, both in terms of language content and of kinds of student behaviour expected. In order to characterise levels, we need tests.

This is casting tests in a different role. I could characte-

rise one difference between structural and communicative approaches as follows: in a structural approach, the syllabus lays down the path to follow, and the teacher attempts to make the students conform to the syllabus; in a communicative approach the teacher attempts as far as possible to make the syllabus conform to the students -- i.e., to reflect their own natural path to language proficiency. Tests help characterise this natural path, to the extent that it can be generalised to the student group as a whole. Good tests will allow us to place students at a point along the path; to group students of similar level, and to check how fast students are making progress.

The information we can hope to get from a discrete-item, paper test of the kind discussed here is clearly not sufficient to characterise levels on its own, but I believe it may be relevant, as long as the items are well designed. We may expect to be able to test active and passive knowledge of spelling, vocabulary idiom and sentence grammar. What we can test depends on the kind of test items that are included.

THE TYPES OF TEST ITEM INCLUDED

As Alderson [1986] has pointed out: '...paradoxically, the computer has been a force for conservatism and lack of innovation in testing techniques: technological innovation has discouraged innovations in content validity.' He is thinking of the preference for simple computer-markable techniques such as multiple choice or true-false questions, which of course cannot test students' active ability to produce or manipulate words or structures. With this criticism in mind I have tried to include a wider range of item types, within the following constraints. Items must: (a) be machine markable; (b) allow a finite, predictable number of correct responses; (c) be short; (d) be unrelated to each other; (e) be markable right or wrong.

Constraint (a) follows from the inclusion in the system of a computer-adaptive testing module in which students answer directly questions set by the computer. Thus the computer must contain all possible correct answers for each item. If the bank

were used only to generate paper tests, this would of course be unnecessary.

Constraint (b) follows from constraint (a): the computer cannot hold more than a limited number of correct responses. Therefore items must be carefully designed to elicit only the expected response.

Constraint (c) follows from the limited storage space available on two floppy discs. To fit 2000 items in the bank each item can be no more than about 250 characters long, including all answers.

Constraint (d) follows from the need to rank items according to their unique difficulty rating. It rules out the use of, e.g., cloze tests with several items within one connected text.

Constraint (e) follows from the statistical analysis adopted. It is another reason for trying to keep items simple.

Five types of test item are included in the present system: (1) Multiple choice; (2) Sentence transformation (3) Use the word (4) Sentence expansion; (5) Gap fill.

Apart from type (1), all are types of guided sentence completion exercise, requiring the candidate to write a phrase or sentence in response to a prompt, the prompt being designed to limit the range of things the candidate might appropriately write.

It is planned that a large part of items should be written by teachers on the basis of errors observed in their students. If this proves practical, it will be further grounds for believing that the graded item bank constitutes a genuine characterisation of our Polish students' path to proficiency in English.

USING THE SYSTEM

The user selects from a menu of options. I shall discuss these options in the order they would be used.

Entering new items. The user selects from a menu the item type, and then screen prompts guide him to enter the prompt and answer(s). There is an editing mode allowing the user to change or delete items.

Entering variant answers for the sentence-completion item ty-

pes is done reasonably easily and economically, using a transition network formalism.

For each item the user can select one or more tags from a choice of 36, e.g., 'lexis', 'idiom', 'tense', 'perfect', and 'comparison'. These can be used later when generating tests to specification.

Compiling, printing and administering a Trial Test. The user selects the Trial Test Print option and specifies how many items from the trial item bank are to be included. The computer then prints a hard copy of the complete test paper, together with introduction and example questions.

Inputting results. The papers are duplicated and the test administered with a representative sample -- say, 100+students from all levels taught. The papers are marked by hand. It will then with certainty be necessary to edit the range of possible answers held in the trial item bank, to allow those acceptable ones which the item writer failed to predict (the system at this stage will allow answers to be edited, but not prompts, as this would alter the character of the item). Editing the possible answers on the basis of the trial sample's responses allows some confidence that the computer- adaptive test module will indeed assess candidates reasonably fairly. The user selects the Enter Results option and is guided to input the results on each item for each candidate. When all results have been entered the program proceeds to analysis.

Analysis -- Latent Trait Theory. The analysis is based on a statistical approach known as latent trait theory or item response theory. I have taken the method, as indeed, many of these ideas for item banking, from the work of Grant Henning of UCLA. Henning [1987] describes a logistic model known as the Rasch Model.

Unlike classical test theory, the Rasch Model can not only grade persons and items for ability and difficulty, but also judge the probability of their response patterns, thus identifying items which discriminate poorly, and persons who respond erratically.

Figure 2 (adapted from Henning, 1987) shows two persons and four items, located at points along the latent continuum. Person

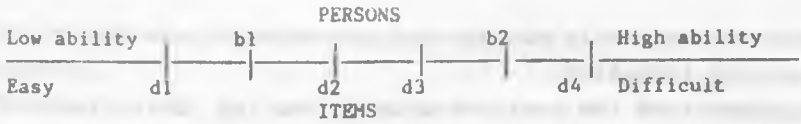

Fig. 2

b2 is more able than person b1, and is also above the level of items d1, d2 and d3. Probably Person b2 would pass items d1, d2 and d3, and fail item d4. The probability of person b2 passing item d1 is greater, however, than the probability of passing item d3 or failing item d4 because these two items are closer to the actual ability of person b2. After estimating the person abilities and item difficulties, the exact probability of particular responses can be calculated, and thus a measure of goodness of fit obtained.

Henning claims many potential advantages of the latent trait model, including the following:

(a) Sample-free item calibration; i.e., we can derive an item difficulty scale which is independent of the particular sample the items were trialled on.

(b) Test-free person measurement; i.e., we can administer different tests to different people and compare the results directly, without having previously equated the two tests. This is possible as long as both tests include a small *link* of common items.

(c) Identification of guessers and other deviant respondents; i.e., people whose response pattern exceeds some level of improbability.

(d) Test equating facility. The model seems to offer a way of linking different tests to each other without the need to give all forms of tests in their entirety to a common or demonstrably equivalent sample of examinees. The group of common linking items provides the key to equating different tests administered to different samples of individuals drawn from the same general population.

(e) Item banking facility. Once items have been calibrated, they can be stored in an item bank according to a common metric of difficulty. The bank can be used to construct tests of known

reliability and validity without the need for further trialling.

Henning describes an approximative method for hand-working, called PROX (Wright and Stone, 1979). During analysis a number of well-fitting items of median difficulty are selected as the link for the next trial test. These items will be included in the next trial test, and their mean difficulty rating on the two tests compared. The difference in mean difficulty is the difference in the ability level overall of the two sample groups, and constitutes the *translation constant* -- the adjustment that must be made to the difficulty ratings of the new test items before they are transferred to the bank.

A list of examinee ability ratings can be printed out, together with measures of goodness of fit. Items which perform erratically are rejected, and the others are transferred to the main item bank, together with measures of their difficulty and goodness of fit. The items in the bank are arranged in order of difficulty.

Compiling profiled tests from the bank. When this process has been repeated several times and the bank contains sufficient items, it can be used to generale tests to specification. Parameters offered include:

(a) the lower and higher difficulty limits, i.e., the level and range covered;

(b) the type of item to be included: the user can specify, e.g., Multiple Choice only, or any combination of item types;

(c) the language areas to be included: this is possible as each item has been tagged with one or more labels;

(d) history of use: items can be excluded if previously used in specified tests.

After the bank has been searched the number of matching items is reported. If too few, the user can change the parameters, or if too many the best items (i.e., best fitting) will be selected to the required number. The test is then printed to paper.

The computer-adaptive test. Computer-adaptive testing is an attractive idea where small numbers of candidates need to be assessed quickly, e.g., when late applicants seek admission to existing groups. The examinee sits at the keyboard and types an

answer to the first item, which the computer selects from the middle range of difficulty. If the examinee responds correctly a more difficult item is selected, if wrongly an easier item, and this process repeats until the examinee´s level is established within prescribed levels of probability. If an acceptable level of probability is not reached within a certain number of items it can be concluded that the examinee is responding erratically, and must be assessed by some other means. The algorithm for this selection-assessment process I have also taken from Henning.

INITIAL RESULTS

At present most of the software for the system exists in an implementation for BBC computer (B plus second processor, or Master). I hope to develop an IBM PC version. The software still needs some refinement and completion. As of June 1988, three trial tests had been conducted, from which 100 items survived the analysis and were transferred to the main bank. This is not many, but it is already enough for the CAT module to draw on and produce assessments. One test has been compiled from items in the bank. It was conceived as a placement test, taking items from almost the whole range of difficulty, and was used as part of the entrance exam for new course participants administered in May 1988. No serious statistical study of the results produced by these tests has yet been undertaken at the time of this writing.

It is thus far too early to begin to assess the system and its utility, or to pass judgement on the bold claims made for latent trait theory. One can say that the analysis identifies and rejects a reasonable-looking number of items as erratic performes, and that it correctly identifies erratic performance in examinees -- i.e., people who do unexpectedly well on some difficult items or fail on relatively easy ones. I am not sure how much significance to attach to the latter phenomenon; after all, language knowledge cannot be expected to be very homogenous even within a group of similar learners. It will be very interesting, when we have enough data, to correlate performance on different profiled tests, and on paper tests as opposed to the computer-adaptive module.