ResearchNotes

UNIVERSITY of CAMBRIDGE ESOL Examinations
English for Speakers of Other Languages

EXTRACT FROM RESEARCH NOTES ISSUE 3 / NOVEMBER 2000 / PAGES 4–6 **3/02**

The use of Rasch Partial Credit Analysis in test development

SIMON BEESTON, EFL VALIDATION MANAGER, UCLES

Previous articles in this series have reported on the processes that UCLES EFL use to ensure that the items it uses in its tests are of the highest possible quality. In this article I am going to look in some detail at a particular type of statistical analysis that reveals the internal functioning of items such as the sentence transformation tasks that are used in the First Certificate in English Use of English paper.

A sentence transformation requires students to rewrite a sentence to produce a new sentence which means the same as the original. Part of the new sentence is provided, along with a key word which must be used to complete the new sentence. An example of this type of item is given below.

'Do you know if it will take a long time to develop the film?' the customer asked me.

would

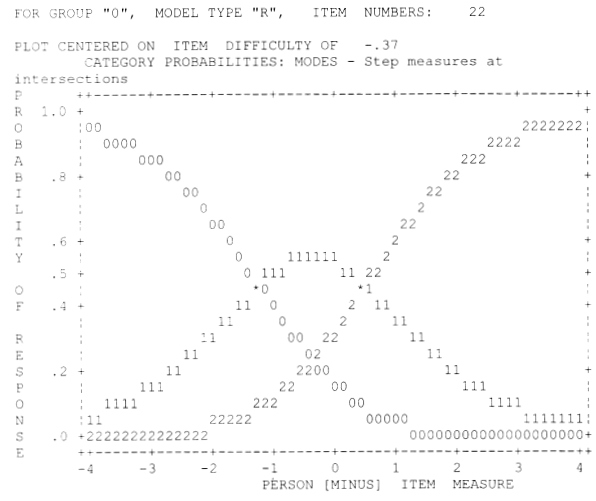
The customer asked me if I knew how ..... to develop the film.

The phrase required to complete the sentence is 'if I knew how *long it would take* to develop...' although there may be other possible variations that would be marked correctly. Students successfully completing the sentence would be given two marks; however, if they completed it partially correctly they could still be awarded one mark. For example, an answer including only 'how long' would be awarded one mark for correctly collocating how with long. It is therefore possible to achieve a mark of 0, 1 or 2 depending on the degree of correctness. Items such as these are called scalar items because students can achieve a mark along a scale, albeit a rather short one.

Apart from the issues of item difficulty and item discrimination, it is also important for this type of item that the scale functions properly; that is to say, that for this item, there is a realistic chance of getting 0 or 1 or 2. To determine whether or not this is the case, we use the Rasch Partial Credit model to evaluate how well the scale is working.

Rasch analysis belongs to a branch of test statistics known collectively as Latent Trait Theory. Latent Trait Theory, a term generally used interchangeably with Item Response Theory, arose from dissatisfaction with traditional approaches to educational or psychological measurement. The name 'latent trait' captures the idea that the underlying scores people obtain in a test correspond to a certain amount of ability, not directly observable, but inferable from observing their performance. Furthermore, each test task has a certain amount of difficulty which can be measured relative to the difficulty of other items in the test. Ability and difficulty are mutually-defining terms which allow persons and test items to be ranged along a single unidimensional continuum. In the case of scalar items such as the sentence transformation reviewed here, the software (Bigsteps; Wright and Linacre, 1994) used by UCLES produces plots showing the probability of each score category being used in relation to the range of ability of the candidates (from low to high ability). This appears as three overlapping distributions (0, 1 and 2) which for a well-functioning, item would produce a roughly symmetrical pattern like the one shown below. The vertical axis provides a measure of probability, the horizontal axis shows increasing ability from left to right. The 0s, 1s and 2s that form the peaks and troughs of the graph show the likelihood of a candidate at a particular ability getting a particular score. As the ability increases, so the likelihood of a higher score increases/ Plot 1 below shows that mid-range candidates have a high probability of scoring 1 and that in fact, the scale works well.

Plot 1



However, one might predict that if the required phrase were something like 'put up with it' where one mark was awarded for 'put up' and another for 'with it', then students would be likely to get all of the phrase or none of it. In such a situation candidates would be more likely to score 0 if they do not know the phrase or 2 if they know all of it. Of course, UCLES item writers endeavour not to produce sentence transformations that function like this but

because we pretest all of our material before it is used in a live test, we are able to identify those sentence transformations which do not work as well as they should.

The next example shows a sentence transformation that had been through careful editing but pretesting analysis revealed that the task was not suitable for a live examination.

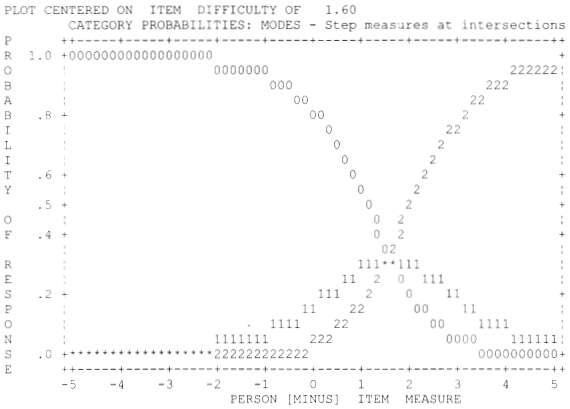
Stephanie lives so far from the school that she has to catch a bus at 6.30 every morning.

way

Stephanie lives ..... from the school that she has to catch a bus at 5.30 every morning.

The required answer is, of course, '*such a long way*' but pretesting revealed that candidates either knew all of the answer or none of it. Compare the distributions of probable scores in Plot 2 below with the earlier well-functioning plot.

Plot 2



As can be seen from this second plot of probable scores, the likelihood of scoring 1 is always lower than that of scoring either 0 or 2. Instead of being a scalar item, it effectively becomes a dichotomous item that is double weighted.

UCLES EFL use a number of different quality control stages to ensure that all material used in live examinations is of the highest possible quality. Using pretesting statistics to evaluate how scalar items are functioning is one such quality control stage in a process that is designed to ensure our examinations are fair, reliable and useful for all our candidates.

References:

Wright,, B. D. & Linacre, J. M (1994): *BIGSTEPS*, Mesa Press

This is an extract from *Research Notes*, published quarterly by University of Cambridge ESOL Examinations.
Previous issues are available on-line from www.CambridgeESOL.org
© UCLES 2003 – this article may not be reproduced without the written permission of the copyright holder.