

BULATS: A case study comparing computer based and paper-and-pencil tests

NEIL JONES, RESEARCH CO-ORDINATOR, UCLES EFL

Introduction

The growth of computer based testing

Computer based (CB) testing is a relatively recent development in UCLES EFL and in many ways is still in a developmental stage. Compared with the large candidatures for the major paper-and-pencil (P&P) exams the current market for CB products is generally associated with low-stakes testing: they are not certificated in the same way as the main suite exams, the conditions in which they are administered are not supervised by UCLES, and they are shorter.

However, in the future this situation will change. The administration of certificated exams, probably online, is a possibility and is an area of current research. A CB version of IELTS has been trialled, and will be made available as an alternative format in 2001.

Current CB products produced by UCLES EFL, in partnership with ALTE members, include:

- BULATS, a CB alternative to the P&P BULATS (Business Language Testing Service), available in English and French;
- Linguaskill, a computer adaptive test (CAT) with a business focus, developed for Manpower Europe and now available in English, French, German, Spanish and Dutch;
- Placement tests under development for the British Council and OUP.

Comparing CB and P&P tests

All UCLES EFL tests and exams provide results which can be interpreted in terms of ALTE levels. Thus there is a general requirement to ensure that in terms of level, there is comparability across all products, CB and P&P. This is also true of different language versions of multilingual CATs like Linguaskill. A major area of research, which is particularly important for establishing this kind of comparability, is the development and use of 'can-do' statements to provide a basis for defining levels in functional terms. An update on this project was reported in *Research Notes 2*.

Every test contains measurement error, and is subject to practical constraints such as length, range of skills tested, etc. In comparing CB and P&P formats, it is important to distinguish general issues of reliability and test relatedness, which affect any comparison between tests, from specific issues relating to the testing format.

Specific issues in the comparison of CB and P&P formats include:

1. The difficulty of particular task types and of individual items;
2. The overall level of performance, and the spread of scores;
3. The impact of such features of CB test administration as time limits, enabling or disabling review of earlier responses, etc;
4. The effect of such test-taker features as gender, age, or familiarity with computers, both individually and when grouped e.g. by country of origin, professional background etc.

These are relevant to the comparison of CB and P&P formats of a linear test such as IELTS.

Additionally, where the CB test is adaptive, (e.g. Linguaskill, BULATS, and the OUP and British Council Placement Tests) the following issues arise:

1. The effect of an adaptive mode of administration on test reliability, discrimination and the effective scale length of the CAT format;
2. The effect of guessing in the P&P format.

This paper focuses on a particular project which was recently completed: a comparison of the CB and P&P forms of BULATS, which addresses several of the issues listed above.

Item banking: the basis of comparability

It is important to understand that when we compare scores between CB and P&P formats we generally do not mean raw scores. Most current CB products are adaptive tests. In such a test candidates will tend to achieve roughly similar proportion-correct scores. But clearly a candidate who scores 60% on a set of difficult items has demonstrated more ability than the candidate who scores 60% on an easy set of items. The scores we are talking about are actually ability estimates derived from a latent trait (Rasch) analysis (see Simon Beeston's article on page 4 for an introduction to Rasch measurement). Similarly, the raw scores on the P&P version are Rasch-analysed to derive ability estimates. It is these which we can compare.

To estimate ability using Rasch techniques we must first know the difficulty of each item in the test, and a basic condition for constructing comparable tests is that the items used in both should be taken from a pool, or item bank, of items which have been calibrated (their difficulty estimated) on the same scale. UCLES EFL has for some years been using item banking techniques in the routine test construction cycle, so that generally when items are made available for use in a CB test their difficulty is known with some precision.

The BULATS comparability study

Earlier this year, 85 learners of English agreed to do a CB and P&P version of the BULATS test of English. They also completed a questionnaire.

Findings from the questionnaire

Candidates were asked to say:

1. How difficult they found the two forms of test;
2. Whether they found the two forms of test to be of appropriate length;
3. Whether they liked using computers;
4. Which version of the test they liked best;
5. Whether they considered themselves good at using computers.

The questionnaire produced some interesting findings, but no evidence that personal attitudes to computers affected performance on the test.

Most people said they liked using PCs. More than half preferred the CB version, and there was a clear tendency for people who preferred the P&P version to say that they found this version easier than the CB version. There was also a small tendency for people who claimed not to be good at using computers to say they found the P&P test easy, but the CB test hard. However, there was no relation between any of these statements and the final score in either form of the test.

These findings suggest that for this group of subjects, who were studying in Cambridge when they took the tests, there was no effect on scores connected with computer familiarity, like or dislike. This in turn suggests that typical BULATS candidates would most probably not be disadvantaged or advantaged by such factors.

Reliability of each test

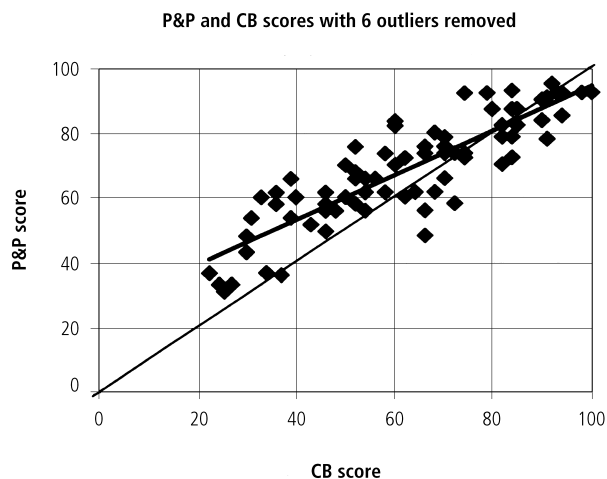
For the P&P test, classical Alpha and Rasch estimates of reliability were .93 and .92 respectively. An average reliability was estimated for the CB tests of .94. Thus both these tests show good reliability for this sample of respondents.

These reliability estimates are based on internal consistency estimates. It would be useful to have coefficients of stability from test-retest data. These could be directly compared with the correlations found between CB and P&P formats, and would thus indicate whether differences in test format have a significant effect on correlations. However, in the absence of test-retest data we can use the square of the alpha reliability to model the correlation between two sittings of the test. This gives .88 for the CB format and .87 for P&P.

Correlation between CB and P&P scores

Figure 1 shows a scatterplot of the CB and P&P scores. The correlation before outliers are removed is .77. With 6 outlying cases removed it is .86. Inevitably the experimental conditions, where the two tests were completed one after the other, produce variations in performance due to fatigue, inattention etc. Removing a small number of outlying cases is sufficient to produce an actual correlation between test formats which is similar to the modelled

Figure 1 : BULATS CB and P&P ability scores compared



test-retest reliability for each test format taken separately (as presented above). While actual test-retest data will allow us to settle this question with more confidence, it appears that the effect of test format on the correlation of test results was minimal for this group of respondents.

Overall level and spread of scores

Figure 1 indicates (from the way the points are distributed along the identity line) that there is good agreement in overall level between the scores obtained on the two formats. However, the spread of scores is clearly narrower for the P&P format, as indicated by the slope of the trend line which has been added. In other words the CB test is slightly more discriminating. The linear trend describes the relationship well: curvilinear trends (e.g. 2 or 3-order polynomials) do not account for significantly more common variance.

Table 1 shows the mean and SD of scores on both formats. The P&P scores are higher overall, and this is mostly caused by lower ability candidates performing better on the P&P version.

Table 1 : Mean and SD of scores on CB and P&P test formats

	CB	P&P
mean	2.80	3.11
SD	1.24	1.15

The narrower spread of scores on the P&P version of a test has been observed previously in other contexts, and is characteristic. The adaptive CB test selects the most appropriate items for each candidate, according to their estimated level. It gives each candidate a chance to show just how high or low their level is. The P&P test is the same for all candidates, and necessarily each item gives slightly less information, because it is of inappropriate level for a proportion of the candidates.

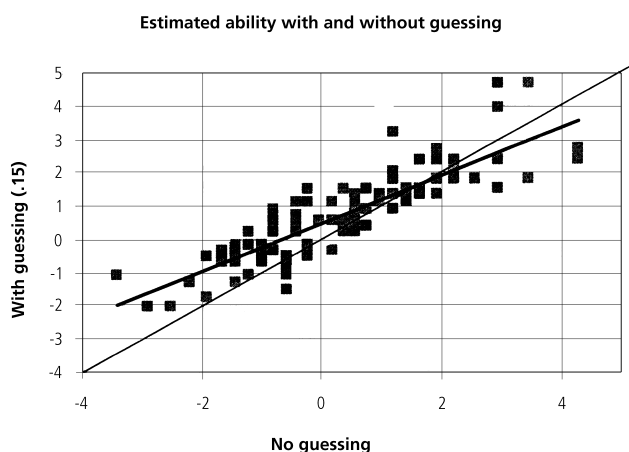
The effect of 'guessing'

It appears that a crucial aspect of this difference between CB and P&P is what is commonly called guessing, although this is better characterised as the contribution of chance in a response to an

item. While there is no systematic benefit from guessing in an adaptive test format, the P&P format does enable candidates who guess to score higher (under normal scoring rules where wrong answers are not penalised).

Guessing is an unfortunate label, because it suggests a distinct, aberrant and relatively rare type of behaviour which occurs only when a respondent finds an item to be wholly too difficult. In fact, what we call guessing is not a distinct type of behaviour, but just the extreme end of a continuum, where the relative contributions of chance and ability are 100% and zero respectively.

Figure 2 : The effect of guessing on ability estimates (from simulated data)



Simulated response data allow us to examine the effect of chance on ability estimates. Figure 2 shows a scatterplot comparing estimates of ability from two artificially-generated datasets. Both sets were generated from the same set of abilities and difficulties. The first used the standard Rasch model; the second used a modified model in which the probability of a correct response tends to be an arbitrary lower limit of 15%.

There is a striking resemblance between this figure and Figure 1 – the comparison of scores on CB and P&P versions of BULATS. What is particularly interesting is that, as with the CB and P&P score comparison, the trend line plotted through the data points is linear. A more complex curvilinear relationship accounts for no more of the common variance. This shows that the effect of chance is not limited to lower-ability candidates, but affects ability estimates proportionately across the whole scale.

Discussion

This paper has not addressed all the issues relevant to the comparison of CB and P&P test formats. The comparability of test content, and the performance of particular task types, have not been treated. However, the findings of the BULATS comparability study described here have contributed significantly to our understanding of how CB and P&P test formats relate, and support a view that it should be practical to develop the two formats for use interchangeably.

Each test format was found to be highly reliable for this group of subjects. The correlation between scores on the two tests was high, and removing just a small number of cases of poor agreement was sufficient to produce a correlation as high as the theoretical (squared alpha) test-retest correlation of each test format taken separately. In other words, there was no evidence of the test format having an important effect on the correlation between two attempts at the test.

The questionnaire also showed no relationship between attitude to computers and test scores on the CB test format, for this group of respondents. Thus on this evidence the two forms of test appear to measure the same thing; however, they clearly measure it on a different scale, as shown by the narrower score range observed for the P&P test format.

The relation between CB and P&P scores was found to be linear. A study conducted on generated response data confirmed that a similar linear relationship could be produced by modelling the effect of chance, or 'guessing', which affects P&P scores much more than CB scores. Thus there is a theoretical explanation for the difference in the observed score distributions, and so it should be possible to equate scores on the two test formats by a suitable linear scaling.

The comparability of CB and P&P formats is practically of great importance. Clearly, decisions on how to report the equivalence of different test formats require consideration of such issues as how high-stakes the test is, who the users of the test are, and whether a simple form of report is practically more useful than a psychometrically rigorous but less transparent one. In the case of BULATS, it seems both reasonable and useful to aim at using a single scale to report scores on both computer-based and paper-based forms of the test.