ResearchNotes
UNIVERSITY of CAMBRIDGE ESOL Examinations
English for Speakers of Other Languages
EXTRACT FROM RESEARCH NOTES ISSUE 5 / JULY 2001 / PAGES 14–16 **5/04**

CB BULATS: Examining the reliability of a computer based test using test-retest method

ARDESHIR GERANPAYEH, RESEARCH AND VALIDATION OFFICER, UCLES EFL

Introduction

The stability of test results over time has been one of the concerns of test designers. One way of demonstrating that stability is by means of test-retest, where a group of candidates sit for the same test twice over a period of time. The Pearson correlation between the scores on the two sittings is called the stability coefficient and is indicative of the reliability of the test. A coefficient of 0.80 or more would generally indicate that the data are reliable enough for practical purposes. Although the stability coefficient is the most appropriate way to show the stability of test results over time, it is not very often reported in language testing literature. This is because it is very difficult to persuade a group of test takers to sit for the same test twice and expect them to take the exam with the same degree of attention on both occasions. This short paper examines the reliability of a computer based test using the test-retest method. The current study follows up the work reported on BULATS by Neil Jones in *Research Notes 3* (November 2000), where a computer-based version of BULATS was compared with the paper-and-pencil version. Jones' study demonstrated that there was a linear relationship between the CB and P&P scores, supporting the view that it should be practical to develop the two formats for use interchangeably. The reliabilities reported for the P&P format and that of the CB were 0.93 and 0.94 respectively and the correlation between the scores on the two tests was 0.86 when six outlying cases were removed. Based on the square of alpha reliability, the study predicted that we would get a correlation of 0.88 between the scores on two sittings of the CB format. The accuracy of such a prediction will be examined in this report by estimating the reliability of a CB test using both the stability coefficient and a Rasch reliability estimate (an internal consistency measure, analogous to Cronbach's Alpha).

The CB BULATS test-retest project

CB BULATS is currently under revision and a new version of the test will be released shortly. The new version, while maintaining the adaptive mode, includes new item types and is relatively longer. As part of the validation exercise, the new version of the test was piloted in Cambridge earlier this year. The main objective of the project was to examine the stability of the new CB BULATS test scores over time using the test retest method. Other issues to be investigated were:

1. The effect of an adaptive mode of administration on test reliability and discrimination, and
2. The effect of test taker features such as L1, gender, age, and familiarity with computers on test scores.

Administration

87 EFL test takers studying at various language schools in Cambridge volunteered to take the new version of CB BULATS twice on the same day with a short break between the two administrations. They were also given a questionnaire to complete. 85 test takers completed the questionnaire. Table 1 demonstrates how the test takers varied with respect to their L1, gender and age.

Table 1 : Test takers grouping by L1, Gender & Age

Grouped by L1 Language

First Language	Frequency	Percent	Cumulative %
Arabic	1	1.18	1.18
Chinese	1	1.18	2.35
Faeroes	1	1.18	3.53
French	3	3.53	7.06
German	8	9.41	16.47
Italian	3	3.53	20.00
Japanese	5	5.88	25.88
Missing	4	4.71	30.59
Portuguese	8	9.41	40.00
Russian	1	1.18	41.18
Slovak	1	1.18	42.35
Spanish	44	51.76	94.12
Turkish	5	5.88	100
Total	**85**	**100**	

Grouped by Gender

Gender	Frequency	Percent	Cumulative %
Female	55	64.71	64.71
Male	30	35.29	100
Total	**85**	**100**	

Grouped by Age

Age Group	Frequency	Percent	Cumulative %
10–16	9	10.59	10.59
17–20	43	50.59	61.18
21–25	16	18.82	80.00
26–34	14	16.47	96.47
35+	3	3.53	100
Total	**85**	**100**	

The candidates' test scores on the two sittings and their responses to the questionnaire were entered into a database for further analysis. For ease of reference, the first administration of the test will be called Test 1 and the second (retest) referred to as Test 2. Reference to results reported in Jones' study will be referred to as Test 3. Test 1 and Test 2 are the new version of CB BULATS, while Test 3 is the current version of the test.

Findings

It is important to mention that test scores on CB BULATS do not refer to raw scores. They are actually ability estimates derived from a latent trait (Rasch) analysis, converted into BULATS scores by means of a scaling procedure. The items in the test and retest were taken from the same item bank with calibrated item difficulties (see Jones, *Research Notes 3* on item banking). The following terminology will be used with reference to the scores: Test Score refers to BULATS test score (0–100), Band Score refers to BULATS band scale (1–5), and Ability level refers to candidate ability as estimated by Rasch model (Logit).

Reliability

The average reliability (Rasch) for each version of the test was estimated as 0.94, and 0.93 for Test 1 and Test 2, respectively. Using the square of this reliability to model the correlation between two sittings of the test, the estimated reliability was 0.87. This figure is very close to the prediction that Jones estimated for the CB BULATS test retest coefficient in his study (0.88). The test scores from Test 1 and Test 2 were correlated to examine how accurate these predictions were. The correlation between the two test scores was 0.89 before any outliers were removed and 0.93 when six outliers were removed. The stability coefficient between Test 1 and Test 2, even before removing the outlying cases, is higher than the value that the square of the alpha reliability predicts. This allows us to be relatively confident about the stability of the new CB BULATS test scores over time.

Figure 1 shows a scatterplot of test-retest scores (with six outliers removed, i.e. replicating the approach used in the previous study). Sitting for a test twice on the same day under experimental conditions will produce variations in performance; however, the high correlation (0.93) achieved between the scores of the candidates on test-retest shows that any such variations were minimal.

Figure 1 : CB BULATS ability scores compared in two sittings

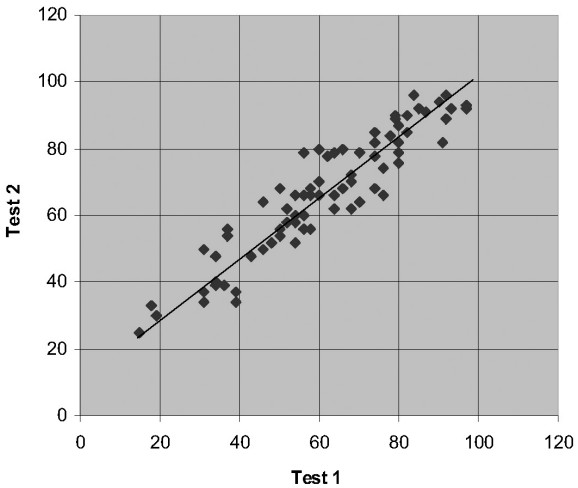

Figure 1 also indicates that there is good agreement in overall level between the scores obtained on the two sittings. The spread of test scores along the identity line shows that the tests are discriminating relatively well; a similar finding was reported by Jones for Test 3. It appears that CB format, in general, can produce more discriminating results. This is due to the adaptive mode of the test, which selects the most appropriate items for each candidate according to their estimated level, providing more information per item and minimising the effect of guessing.

Table 2 reports the mean and SD of band scores of the candidates for the current and new version of CB BULATS. The mean band scores on Test 1 and Test 3 and their variability in scores (SD) are so close that it allows us to conclude that the two populations were similar in terms of their ability. The slight change of band scores in Test 2 is due to the better performance of the test takers on their second attempt. To determine whether the differences in candidates' test scores / bands in Test 1 and Test 2 were significant, t-tests were applied.

Table 2 : Mean and SD of band scores

	Band Scores		
	Test 1	**Test 2**	**Test 3**
Mean	2.78	3.06	2.80
SD	1.32	1.18	1.24

Table 3 illustrates that the candidates scored significantly higher in retest. This improvement in language ability was greatest for lower-level candidates, hence the lower SD of scores observed for Test 2. We will be discussing this in *Final Remarks*.

Table 3 : Results of tests of significance

Variables compared	**t**	**df**	**Sig. (2-tailed)**
Pair 1 Band Score 1 – Band Score 2	-3.396	86	0.001
Pair 2 Test Score1 – Test Score 2 -	4.375	86	0.000

Table 4 compares Rasch reliability estimates for the three CB BULATS tests. The current version (Test 3) and the new version

Table 4 : Test Reliabilities (Rasch)

	Test 1	**Test 2**	**Test 3**
Ability SD	1.29	1.12	1.32
Mean SEM	0.33	0.31	0.33
Separability	3.90	3.61	3.99
Reliability	0.94	0.93	0.94

Test 1 & Test 2 (Test & Retest) = New version of CB BULATS (this study)
Test 3 = Current version of CB BULATS (reported in Jones' study)
Ability SD= Standard Deviation of candidate's ability
SEM =Standard Error of Measurement

(Test 1 & Test 2), despite their differences in format and length, seem to be comparable with respect to their mean standard error of measurement, separability and reliability (Rasch) estimates. The slight decrease in the reliability of Test 2 is due to the better performance of test takers on retest, which resulted in lower variability in scores on their second attempt.

The effect of test taker features on test scores

There are various ways of examining the influence of test taker features such as gender on test results of which Analysis of Covariance (ANCOVA) is one. ANCOVA is a means of reducing systematic bias, as well as within-groups error in the analysis. The aim is to determine whether the independent variable – gender, age, etc. – is indeed having an effect on the dependent variable, i.e. Test 1 scores; we do this by statistically controlling the influence of an extraneous variable such as Test 2 scores (covariate) on the dependent variable. In other words, we attempt to reduce the error variance caused by individual differences.

To examine the effect of test taker features on test scores, a number of *One-Way Analysis of Covariance (ANCOVA)* were conducted on test scores with respect to the information collected through the questionnaire. In each ANCOVA, test score on Test 1 was the dependent variable, test score on Test 2 was the covariate and the feature under investigation was the independent variable. Features investigated were as follows: L1, gender, age, familiarity with computers, frequency of computer use, preference in using CBT and P&P, and suffering from eye strain during the test (Test 1 & Test 2). None of the analyses conducted indicated that there was a main effect ($p>.05$) for the features examined. Thus we can say that test taker features examined in this study seem to have no influence on test scores in CB BULATS. A similar finding was also reported in Jones' study.

Final remarks

This research project followed up the work in Jones' earlier study where, amongst other findings, a linear relationship was reported between the scores of CB and P&P versions of BULATS. The main objective of the present study was to examine the stability of CB BULATS test scores over time and across versions.

We have demonstrated that CB BULATS test scores remain highly stable across versions and over time with a reliability estimate of 0.94 and a stability coefficient of 0.93. We have also shown that familiarity with computers does not seem to advantage/disadvantage CB BULATS candidates. The finding that we have overall higher test-retest agreement for CB-CB (0.93) than for CB-P&P (0.86), however, may indicate that the mode of administration has an effect. This will be addressed in future issues of *Research Notes*.

Finally, we have observed that the candidates scored significantly higher in their second attempt, which might indicate practice effect. Observation of individual cases shows that the variation is greatest in the scores of lower-level candidates. It could be that some of the candidates did not know how or when to key their responses; having done the test once, they had a better sense of what was expected of them. This study did not aim at examining CB practice effect, therefore further speculation does not seem to be warranted at this stage. The practice effect of a CB test can be examined in future research projects.

1 The Business Language Testing Service (BULATS) is a language assessment service specifically for the use of companies and organisations. The service is designed to test the language of employees who need to use a foreign language in their work, and for students and employees on language courses or on professional/business courses where foreign language ability is an important element of the course.

2 See Jones' article in *Research Notes 3* (November 2000), pp. 10-13, for more detailed discussion of computer adaptive testing.

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