7 Statistical characteristics of the test

Two key qualities of an exam are validity and reliability.

Validity relates to the usefulness of a test for a purpose: does it enable well-founded inferences about candidates' ability? Can performance in the test be interpreted in terms of ability to perform in the real world?

Reliability relates to the accuracy of the measurement of the exam: does it rank-order candidates similarly in repeated uses? Can we expect a candidate to achieve the same score in two versions of the same test or in the Computer-based and the Standard tests?

This section presents evidence for the validity and reliability of BULATS.

7.1 How accurately do the BULATS tests measure?

It is important for candidates and exam users to be confident that an examination produces scores that are accurate in that 1) the scores within one test are not significantly different for candidates who are at the same ability and 2) if a candidate sat two versions of the same test (and no increase in candidate ability occurred between the tests) he or she would get the same or nearly the same score on both tests.

Language testers use the concept of *Measurement Error*, for example, to observe this. *Error* does not mean that the test contains mistakes but rather that candidates' scores are not completely consistent across the test or between different versions of the test. Imagine a group of candidates who are all at the same level of language proficiency. If they sat a test they would not all get exactly the same score no matter how accurate or long the test was. This difference in scores could be due to a number of factors such as different levels of motivation or misinterpretation of a question or some candidates meeting questions that tested a particular area of language they were weak at. This difference in scores is an example of Measurement Error.

Reliability

A common way of measuring error and consistency of test scores is to use a correlation coefficient called *Cronbach's Alpha*. This operates by dividing the test into halves and correlates the candidates' scores in one half of the test to their scores in the other half. It then adjusts the correlation to take account of the full number of items in the test as a whole. In theory, reliability coefficients can range from 0 to 1 but in practice we can expect them to be between 0.6 and 0.95, with the higher number indicating a more reliable test.

BULATS agents are requested to return candidate answer sheets from the Standard test to Cambridge ESOL, where they are used to calculate the reliability of the BULATS Standard test. Table 1 opposite shows the reliability (Cronbach's Alpha) for the most recent versions of Standard BULATS based on a random sample of the live BULATS population.

Table 1: Reliability (Cronbach's Alpha) for most recent versions of Standard BULATS, by component and as a whole

Standard test version	Sample size of candidates	Listening reliability	Reading and Language Knowledge reliability	Overall reliability
EN21	520	0.93	0.95	0.97
EN22	468	0.92	0.92	0.96
EN23	959	0.92	0.93	0.96
EN24	1446	0.95	0.94	0.97
EN25	789	0.91	0.92	0.95

The above table shows that the error in candidates' scores due to non-linguistic factors such as candidates misinterpreting a question or losing concentration during a test contributes only a very small part to their overall score. These tests are accurate in placing candidates at the same level in relation to other candidates who sat the test.

However, correlations depend on the rank ordering of candidates: consistent rank ordering is easier to achieve with a group of candidates with a wide range of abilities. Therefore, measures of reliability, such as Cronbach's Alpha, are as much dependent on the spread of ability of the candidate population as the accuracy of the test. This means that direct comparison of reliability across different tests with different populations and ranges of item difficulty can be misleading. In judging the adequacy of the reliability of a test we need to take into account the type of candidates taking the tests and the purpose of the test.

The BULATS Computer-based test is adaptive, which means that candidates receive items appropriate to what the test calculates as the candidate's ability. Therefore different candidates will not be presented with the same items. This means that split-half methods of calculating reliability, such as Cronbach's Alpha, cannot be used. An analogous measure, the Rasch reliability, is used instead; rather than raw scores, this reliability measure uses candidates' ability estimates.

The reliability of the overall test is 0.94, which is very high. At first glance, it may be surprising that the reliability of the Listening sub-section (0.92) is higher than that for the RLK sub-section (0.89) since the Listening section is shorter than the RLK section. However, the Listening section also shows a higher standard deviation of ability estimates which would help to increase the reliability, since reliability improves as the collection of score data becomes more widely spread from the mean and the range increases.

Standard Error of Measurement

Another way of describing the accuracy of a test is in terms of candidates' individual scores and the likely variation in those scores from their real or *true scores*; that is their scores if the test contained no Measurement Error whatsoever. (A *true score* can be defined as the mean score if a candidate were to take the test repeatedly.) This is what the *Standard Error of Measurement* provides.

The transformation of raw scores to BULATS scale scores is non-linear. Therefore the form of SEM that is most meaningful to report is the *Conditional Standard Error of Measurement*. This relates to a particular score in the test. In the case of BULATS, Conditional SEM is most useful when used to estimate the error associated with each band cut-off, that is the lowest score at a band. The conditional SEM will vary slightly according to test version and band cut-off, depending on the precise

difficulty of the items. However, the values reported in the table below for a sample calibration version are typical.

Table 2: Conditional SEM in BULATS Standard scores for a sample calibration version

At band cut-off	Overall SEM	Listening/Reading and Language Knowledge SEM
5	+/-3	+/-4
4	+/-4	+/-5
3	+/-4	+/-5
2	+/-4	+/-5
1	+/-3	+/-4

The table above shows that candidates at Band 1 or 5 are likely to get a score that is within 3 points of their *true score*. They are almost certain to get a BULATS score within 6 points (2 Standard Error Measurements) of their true score. For candidates at Bands 2–4 these numbers are 4 and 8 respectively.

It is always possible that a candidate will be at the borderline between 2 levels, but for the majority of candidates who take the BULATS Standard test the Standard Error Measurements reported above show that they will receive a band that is an accurate reflection of their true ability. It is extremely unlikely for candidates to receive a band that is more than one band higher or lower than their ability warrants.

If we want to compare candidates' scores, either those of two different individuals or the same candidate's performance over time, it is necessary to take into account the Standard Error of Measurement of both scores. This is higher than that for a single score. We can calculate from the table above that a difference of 7 BULATS points between candidates probably indicates a real difference in language ability and candidates with a difference of 14 BULATS points are almost certainly at different language abilities. However, when comparing candidate performance over time it is also necessary to take into account another aspect of Measurement Error known as *Regression to the Mean*. This is a statistical phenomenon whereby candidates who score well below or above the mean will tend to score nearer the mean if they sit the test again regardless of any improvement in language ability. Regression to the Mean is a phenomenon of all tests.

As with reliability, the SEM of the computer-based tests is measured in terms of candidate ability, rather than raw scores. However these can be converted to raw score equivalents.

Table 3: SEM (RAsch) raw score equivalents in CB BULATS Version 6.1 (sample size = 1407)

Overall SEM	Reading and Language Knowledge	Listening
+/-5	+/-6	+/-7

As for the table detailing Standard test SEM, the table above shows that the majority of candidates will receive a band that is an appropriate reflection of their true language proficiency level. It is highly improbable that a candidate will receive a band that is more than one level different from their true language proficiency level.

Reliability in the Speaking and Writing tests

BULATS Speaking and Writing tests contain tasks that are authentic in that they resemble tasks that candidates might be expected to perform in a business environment, such as writing a report or giving a short presentation. These types of tasks require assessment by trained examiners. Reliability in BULATS Speaking and Writing tests is centred on the need to ensure that examiners mark consistently over time (intra-rater reliability) and with other examiners (inter-marker reliability). This is maintained in the examiner certification and training process as detailed below.

- The Writing test assesses writing skills in relation to the workplace, and takes 45 minutes. The test is assessed by two trained language specialists.
- Writing examiners undergo a rigorous training programme in order to qualify and are required to re-certify every two years. In addition, sample scripts are regularly monitored by examiner monitors to ensure that examiners in the field are marking to accepted standards.

- The Speaking test assesses speaking skills in the workplace. The test is recorded and assessed by the examiner conducting the test and then by a second trained examiner. Provision is made for a third examiner to assess the interview in cases where examiner 1 and examiner 2 differ in their grading by more than two sub-bands.
- Oral examiners undergo a rigorous training programme in order to qualify and are required to re-certify every two years. In addition, sample interviews are regularly monitored by examiner monitors to ensure that examiners in the field are marking to accepted standards.

7.2 Are different versions of the tests equivalent?

Organisations often use BULATS over an extended period of time. Therefore they are likely to use more than one version of the Standard or Computer-based test. It is essential that candidates and exam users are confident that different versions of the test are equivalent in that they produce scores and bands that are at the same level of language proficiency.

The equivalence of different versions of the tests is promoted by the Examination Production Cycle, where item writers are trained and items are trialled and checked for suitability and difficulty. This process is explained in detail in section 8.4. In particular, different versions of Standard BULATS are equated to each other in a two-stage process. New items are pretested together with anchor items of known difficulty. This allows the difficulty of the new items to be calculated. Items accepted for live use are then calibrated in a second exercise, where new items and anchors are administered in live test conditions to a specified sample of candidates. Care is taken to ensure that this sample is representative of the BULATS candidate population in terms of first language background and language level. This allows us to check to see if any of the items show bias, that is that they are particularly difficult to a specific group of candidates for non-linguistic reasons. Any such items are excluded from the final test. Items which appear adequate then enter an item bank, from which new BULATS versions are constructed that conform to set targets of Item Difficulty and Discrimination. Item Difficulty is a measure of the likelihood of a candidate of a fixed ability to answer the item correctly. Discrimination is a measure of the capacity of an item to distinguish or discriminate between weak and strong candidates. Item banking also allows tests to be constructed that contain items that test a representative sample of the grammatical structures, functions and topics associated with English in a business environment.

For each new BULATS version transformation tables are produced which convert raw scores to BULATS standardised scores and BULATS bands. These tables are produced for Reading and Language Knowledge, and Listening, and for all items separately, so as to provide the component BULATS score and overall BULATS score and band.

For the BULATS Computer-based test, the equivalence of different versions is maintained through the Standard test, from which calibrated items are chosen.

■ 7.3 Are the Computer-based test and the Standard test equivalent?

Some organisations use both the Computer-based test and the Standard test. In these situations it is essential that exam users are confident that both tests produce scores that are comparable.

Whenever a new Computer-based test is designed its items are taken from calibrated items in the Standard test. To investigate the effect of the mode of the test (computer or paper based) a sample of candidates from a number of different languages are requested to take both the Computer-based test and a Standard test. Their results are correlated and bands and scores in each test are compared to ensure that candidates receive similar scores in the Computer-based and Standard tests taking into account the SEM for both tests.

Below is a table showing the correlations of scores in recent computer-based and standard versions of the test. The measurement error of the tests, as discussed earlier, underestimates this correlation. This is taken into account and is known as *Correction for Attenuation*. This table shows that the mode of the test (paper or computer-based) has, in most cases, little effect on a candidate's band and overall score. However, for the minority of candidates who are uncomfortable using a computer we cannot expect their scores to be the same in each mode.

	Correlation (corrected for attenuation)
Band score	0.86
Overall BULATS score	0.95
Sample size	62

■ 7.4 On-going validation of BULATS

Cambridge ESOL places emphasis on the maintenance of quality by regular monitoring of candidate and examination performance. Re-appraisal and, where necessary, revision to ensure that examinations provide the most accurate, fair and useful means of assessment are key strengths of the organisation. This work is supported by the Research and Validation Group at Cambridge ESOL; the largest dedicated research team of any UK-based provider of English language assessment.

Currently a number of projects are in progress related to BULATS and Business English. Volume 17 in the Studies in Language Testing Series: Issues in Testing Business English by Barry O'Sullivan, deals with the recent revision of the Business English Certificate (BEC) examinations and outlines Cambridge ESOL's understanding of the Business English Construct and model of communicative ability.

On occasion, organisations and candidates are requested to help in providing data and feedback for research projects; their co-operation is welcomed.

More information on Cambridge ESOL and its research and validation work can be found on the Cambridge ESOL website: **www.CambridgeESOL.org**

Cambridge ESOL produces Research Notes, a quarterly journal dealing with current issues in language testing and Cambridge ESOL examinations. These can be accessed on the website and a search for articles related to BULATS or other themes made. A selection of recent articles on BULATS and Business English is given below.

BULATS: A case study comparing computer based and paper-and-pencil tests
Neil Jones Research Notes Issue 3 (November 2000)

CB BULATS: Examining the reliability of a computer based test using test-retest method
Ardeshir Geranpayeh Research Notes Issue 5 (July 2001)

Revising the BULATS Standard Test
Ed Hackett Research Notes Issue 8 (May 2002)

Some theoretical perspectives on testing language for business
Barry O'Sullivan Research Notes Issue 8 (May 2002)

Analysing domain-specific lexical categories: evidence from the BEC written corpus
David Horner and Peter Strutt Research Notes Issue 15 (February 2004)

Using simulation to inform item bank construction for the BULATS computer adaptive test
Louise Maycock Research Notes Issue 27 (February 2007)