The *Password* Test – Design, Development And Reliability

Dr Tony Green, Reader in Language Assessment, CRELLA, University of Bedfordshire

Introduction

English Language Testing Ltd (ELT), the creators of the ***Password*** tests and the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire, its academic designers and managers are committed to on-going research into the ***Password*** test's performance**.** This study into ***Password's*** reliability and accuracy forms a part of that process and is based on data from over 5,000 representative test takers. The evidence is that the ***Password*** test is an extremely reliable test, discriminating effectively from the A2 into the C1 level of the Common European Framework (approximately IELTS 3.5 to IELTS 6.5) and so confirming that ***Password*** is a valuable tool for its intended assessment, counselling, screening and placement purposes.

The *Password* Test

Password was launched in 2008 by ELT, based in London, United Kingdom, whose shareholders include the University of the Arts, London. The test was designed by the CRELLA with input from a steering group that included representatives of the University of Southampton, the University of Reading and the University of the Arts London, with wider consultation across the Higher Education sector.

Test Purpose

Password is intended to

- assess language knowledge – knowledge of grammar and vocabulary in context – rather than language skills – reading, listening, writing, speaking
- discriminate most effectively from the A2 into the C1 level of the Common European Framework, or from approximately IELTS 3.5 to IELTS 6.5
- indicate the amount of English language instruction that is required before learners will be ready for admission to a university degree level academic course
- indicate the amount of English language instruction that is required before learners will be ready for a test involving extensive text-based reception and production, such as the International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL) or Pearson Test of English (PTE) Academic
- inform decisions on placing learners into the most suitable class for their level of language ability
- screen students joining university degree level academic courses to identify those in need of additional English language support (in-sessional English)

In higher education (HE) institutions where the medium of instruction is English, proficiency in English is a fundamental precondition for academic success. Most international students wishing to access and gain maximum benefit from English-medium academic courses will require preparation both in relation to their language abilities and in terms of the academic culture of the receiving institution.

The ***Password*** test responds to the general need for a quick and inexpensive, but accurate indication of a learner's level of English language proficiency and the distance learners may need to cover in their language learning in order to reach an adequate standard for academic study with little or no language support. While comprehensive skills-based tests are suitable for those who are already equipped to enter English medium academic courses or are very close to this level, many students will need extensive language study before they approach the required standard. For these learners in particular, ***Password*** is a more suitable assessment.

Password Test Takers

The majority of ***Password*** test takers are young, educated non-native English speaking adults between the ages of 17 and 25 from a wide range of linguistic, cultural and educational backgrounds. Most will already have experienced several years of formal instruction in English as a foreign Language, but may not have had previous experience of life in an English speaking country or of hearing English spoken around them.

Using *Password*

Unlike many other English language tests used by HE institutions and related organisations, ***Password*** is securely administered and constantly monitored for quality. The test is based on an extensive bank of material that is regularly updated and monitored statistically to ensure consistency of standards. In this way, each test taker receives a unique selection of items, but can be located on the same measurement scale as all other ***Password*** test takers.

Validity

Validity represents the extent to which the interpretation of test scores is justified by evidence. Tests can be more valid for one purpose than for another, but users will always need to consider the extent to which a test provides relevant information that will help them to arrive at well-informed decisions. Evidence for validity may include the rationale for the design of the test and the measurement qualities of the test questions. Does the test cover the areas of knowledge, skills and ability that are of interest to the test user? Is the test capable of providing consistent and meaningful results?

Rationale For The *Password* Test Design

This section briefly explains why the ***Password*** test focuses on key areas of learners' knowledge of grammar and vocabulary.

Grammar And Vocabulary Are Powerful Indicators Of Overall Language Ability.

It has long been acknowledged that tests of grammar and vocabulary knowledge can provide a useful indication of a learner's general language abilities and of their performance on skills based test components – particularly reading and writing (Weir 1983, Read 2000, Hughes 2003, Purpura 2004, Shiotsu and Weir 2007; Hawkey 2009). Across tests that include components addressing grammar and vocabulary together with skills-based sections, the highest correlations between individual test parts and the overall scores are generally those for lexico-grammatical components such as the Use of English papers found in Cambridge examinations (Hawkey 2009) or the Structure and Written Expression component of the paper based TOEFL test (see for example Educational Testing Service 1997). Grammar and vocabulary components also tend to be the most efficient and reliable sections of a test. They are less susceptible to measurement error than other test sections and so provide more consistent scores.

In fact, the relationship between lexico-grammatical measures and overall ratings of language abilities is so strong that grammar tests are often used by researchers as indicators of general language proficiency (see for example Purpura 1999) and it was argued during the 1970s and 1980s that they were sufficient for the full range of language testing purposes (see Oller, 1979). Indeed, after a comprehensive four year multi faceted test development programme, the high correlations

found between the grammar section of the Test in English for Academic Purposes (TEAP, now TEEP) led Weir (1983, p.521) to conclude that, 'the test of grammar might be a sufficient indicator on its own of a student's ability to cope with the language demands made on students by English medium study'. Similarly, Alderson (1993) notes that the pilot grammar component of IELTS correlated so highly with other components of the test that a distinct grammar component was felt to be unnecessary in the operational test. For both TEEP and IELTS the use of skills based components was favoured over grammar because the test developers wanted to encourage learners to develop their skills in preparing for the test. This is not a concern for ***Password*** as learners will take skills based courses after taking the test.

Taken together with their ease of administration and scoring, the benefits of well designed grammar and vocabulary tests make them very attractive options for placement (Green and Weir, 2002). However, we believe that there are further convincing reasons to favour their use in the specific context of ***Password***.

Grammar And Vocabulary Are Fundamental To All Language Use; Especially For Academic Purposes And Especially For Lower Level Language Learners

Assessment of subject knowledge in academic contexts depends predominantly on academic writing – essays, reports, dissertations and theses – based in extensive reading (see Weir et al. 2009) – we have seen that these skills are the most closely linked to performance on tests of grammar and vocabulary. Even presentations and seminar papers may consist largely of written work presented orally.

Successful academic writing requires accurate use of language both at the level of the phrase and sentence and in the organisation of extended discourse. Research suggests that the development of discourse level skills requires a good level of lexico-grammatical knowledge (Shaw and Weir 2007, Khalifa and Weir 2009) while discourse is rightly a focus for advanced EAP courses. In other words, learners who are able to use a wide range of structures and a good command of vocabulary are likely to benefit most from instruction in discourse level skills and can build their awareness of academic register. Those who are not able to form sentences accurately are unlikely to able to organise their ideas effectively and with sensitivity at the level of the text.

Grammar And Vocabulary Are Common Features Of All Language Learning

When learners arrive to take a language course, they will often be coming from a wide range of educational contexts. As a result, at course entry listening tends to be an unstable skill (Jordan 1997). Some learners will be arriving from locations where they have had not heard English spoken and will need time to adjust themselves to the sounds of English, others may be continuing to study or may have recently spent time in an English speaking country and so have already passed through such an adjustment. This process of 'tuning-in' is usually relatively rapid and over the course of a few weeks learners with a sound grasp of grammar and vocabulary can make very quick progress with listening comprehension while those with longer exposure, but less language knowledge will struggle to improve their comprehension. Tests of listening given at course entry will therefore lack accurate predictive power. We believe that in this context, it is better to consider the relatively stable knowledge of grammar and vocabulary as a basis for placement than to attempt to combine these with measures of listening ability.

Tests of grammar and vocabulary are common in almost every educational system, whatever the favoured method of teaching. This is not true of tests of oral or written production and lack of familiarity with such formats negatively affects performance. This is not a great problem for high stakes tests as it can be assumed that learners are motivated to learn about the test format and practice accordingly (Green 2007). The same assumption cannot be made for a test like ***Password*** which nonetheless needs to be immediately accessible to the full range of test takers. The use of familiar selected response formats ensures that ***Password*** holds no surprises for test takers whatever their background.

Test Development

Password was developed on the basis of Weir's (2005) socio-cognitive framework for test validation. The chief concern is with the processing of language at the word and sentence levels fundamental to both comprehension and production and with the academic social context: we are concerned with the language used in academic textbooks, in student writing and that encountered in the daily lives of students.

Detailed test specifications have been developed to reflect the core language knowledge that students need to acquire before they will be able to cope with understanding and producing academic texts. These specifications are based on a number of sources:

- Research carried out by CRELLA into the nature of academic language use (Weir et al 2009)
- Communicative functions (and their associated grammar and vocabulary) found in popular English text books that are widely used on pathway programmes.
- The Breakthrough, Threshold and Vantage specifications for English describing the A2, B1 and B2 levels of the Common European Framework of Reference for Languages (van Ek and Trim 1991a, 1991b, 2000, Council of Europe 2001).
- Common patterns of error in grammatical structures and vocabulary choice made by pathway learners in their written work.
- Research evidence on the essential grammar and vocabulary needed to support academic study (Weir 1983; Nation 1990). We used corpus based wordlists such as the academic wordlist (Coxhead 2000) and word frequency lists based on the British National Corpus to identify words that learners would need to know in order to access academic texts across disciplines.
- Grammar and vocabulary books designed for learners of English such as Murphy (2004) and McCarthy and O'Dell (2008) and reference books such as Greenbaum and Quirk (1993), Carter and McCarthy (2006) and Schmitt (2000).

Overview Of *Password* Item Development Processes

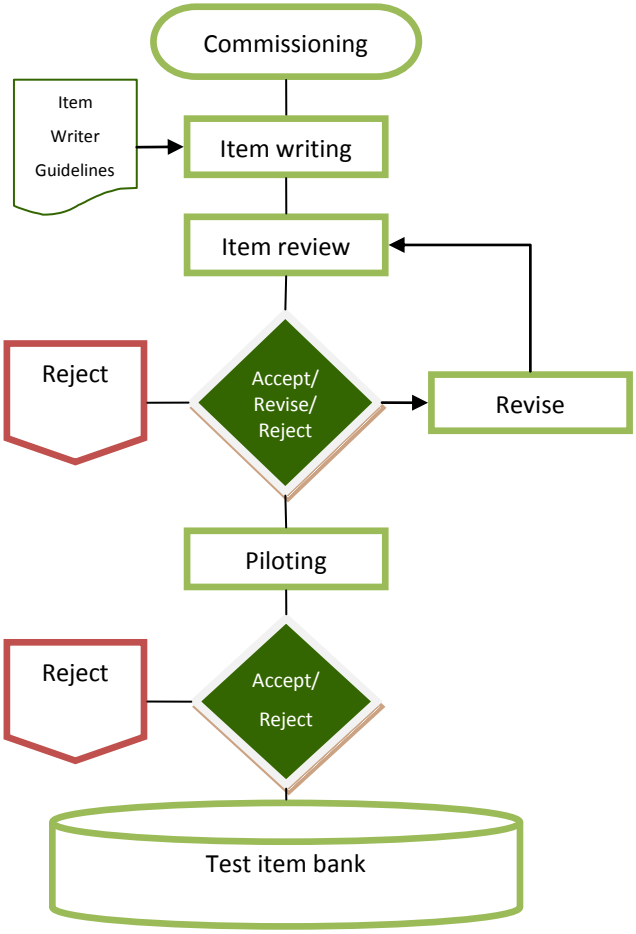

Figure 1 Item development process for *Password*

All ***Password*** questions are written by a suitably qualified team of item writers with a postgraduate qualification in EFL or related field – a Diploma in English Language Teaching (Cambridge ESOL DELTA) or Masters in English Language Teaching or Applied Linguistics – and experience as a teacher of English for Academic Purposes. All item writers are given training in writing items for ***Password*** and follow detailed item writer guidelines (a version of the test specifications that includes detailed information about item characteristics) laid down by the test developers.

The process of generating new test material follows the steps set out in Figure 1, which are explained below.

Commissioning

A regular request is made to the item writers to submit a given number of test items of specified types based on the item writer guidelines. The writers are given a period of three to four weeks to craft and submit a number of items in conformity with these guidelines.

Item Review

A review panel made up of members of CRELLA and ***Password*** staff review all submitted items, considering how well they reflect the guidelines and their suitability in terms of their likely difficulty and of cultural accessibility or sensitivity.

For each item the panel makes one of three recommendations: accept, revise or reject. Accepted items are input to the ***Password*** test delivery system for piloting. Where minor revisions are considered necessary (25% to 35% of cases), these are made by the review panel and the amended items are input for piloting. In 10% to 20% of cases, the items are rejected. Feedback is given by the panel to the item writers on the reasons for revision or rejection.

Piloting

Once they have been accepted, the pilot items are uploaded to the pilot item bank ready for trialling with ***Password*** test takers. A small number of pilot test items are administered alongside the operational test items in each ***Password*** test administration. The test takers' responses to pilot items do not contribute to their official scores, but the results are stored for analysis. In this way, we can be confident that test takers respond to the pilot items in the same way as they do to the operational items.

Once a pilot item has been administered with a sufficient number (at least 250) of test takers it is withdrawn from the pilot bank on the ***Password*** system and analysed statistically. The difficulty of the new item (as measured on the ***Password*** scale) is calculated through a technique known as Rasch analysis. Results on pilot items are compared with results for items of known difficulty from the operational item bank. A small number of pilot items are rejected at this stage either because

they are at a level of difficulty outside the intended range, or because they yield results that are inconsistent with the rest of the test.

Once items have been piloted and their difficulty established, they are promoted to the live test item bank. The performance of items is regularly reviewed to ensure that it continues to be appropriate and items are periodically rested or retired from the operational item bank.

Test Structure

The test consists of 100 selected response items (60 discrete questions), an additional 20 un-scored pilot items (12 discrete questions) are embedded in the test for the purpose of pre-testing. Each correct response is scored as 1 point.

Each test taker completes a short background questionnaire and a can-do self-assessment form before attempting the test. Time spent completing the background questionnaire, can-do self-assessment and on the example items does not count towards the time allotted to the test itself.

Each test section is preceded by instructions and examples explaining the item types in the section.

Test takers are allowed one hour to complete the test (although in practice most complete within 45 minutes).

The test has five sections. Details of the item types in each section are given below.

Test Part	Test Focus	Question format	N°. of questions	N°. of pilot questions	Scoring
Part 1	grammar and vocabulary	3-option multiple-choice – single gap-fill	15	3	15 points
Part 2	grammar and vocabulary	3-option multiple-choice – two or three gaps	12	2	30 points
Part 3	vocabulary: synonymy	15 five-option questions	15	3	15 points
Part 4	vocabulary: collocation	five-option multiple-choice	9	2	9 points
Part 5	grammar and discourse	multiple true-false/ sentence matching	9	2	31 points
Total			60	12	100 points

Sample Questions

Part 1

Getting to College

My teachers helped me so much with applying for colleges. I [ ] have made it through the process without their help!

Part 2.

Glasses

Student A: What [ ] for when I came in?

Student B: My glasses, [ ].

Foot and Mouth Disease

A cat, or a bird that uses infected straw to make a nest, can

[ ] foot-and-mouth disease. Infectious particles can

[ ] be carried by the wind or on our clothes, [accepting] the

disease to move easily from farm to farm.

Part 3.

"Newton's laws are **adequate** for explaining how apples fall from trees."
Which word is most like *adequate*?

- ○ sufficient
- ○ projected
- ○ corresponding
- ○ equated
- ○ dependent

Part 4

"That pile of bricks is **enormous**."
Which word is most often used with *enormous*?

- ○ enormous quality
- ○ enormous amount
- ○ enormous mode
- ○ enormous trace
- ○ enormous code

Part 5

Painting a House

My friends are getting painted their house next week so they are going to stay in a hotel.

- ○ right
- ○ wrong

- **My friends are having their house painted next week** so they are going to stay in a hotel.
  - ○ right
  - ○ wrong

- **Next week my friends' house is being painted** so they are going to stay in a hotel.
  - ○ right
  - ○ wrong

Part 5

The Time

Student A: Why are you so late?

Student B: There was no clock in the room so I didn't know **what time it was**.

- ○ right
- ○ wrong

Student A: Why are you so late?

Student B: There was no clock in the room so I didn't know **what the time was**.

- ○ right
- ○ wrong

Student A: Why are you so late?

Student B: There was no clock in the room so I didn't know **it was what time**.

- ○ right
- ○ wrong

Student A: Why are you so late?

Student B: There was no clock in the room so I didn't know **was what the time**.

- ○ right
- ○ wrong

Scoring And Score Interpretation

Password score	*Password* band	Common European Framework (CEFR)
76 - 100	*Password Plus*	C1 (and above)
69 - 75	6.0	B2
63 - 68	5.5	
57 - 62	5.0	B1
51 - 56	4.5	
45 - 50	4.0	
40 - 44	3.5	A2
35 - 39	3.0	
0 - 34	*Pre-Password*	A1 (and below)

Table 1 *Password* score interpretation

Table 1 shows how ***Password*** scores are reported both as a percentage and as a band score. ***Password*** scores are broadly predictive of outcomes on tests linked to the Common European Framework of Reference (CEFR) (Council of Europe, 2001) and of IELTS scores so that a score of ***Password*** 5.5, for example, would suggest that a learner would be ready to attempt a B2 level test or attend a B2 level language course. Evidence of these relationships can be found in the publication '***Password*** and the CEFR'' which can be downloaded from www.englishlanguagetesting.co.uk.

Test Data For 2010-2011

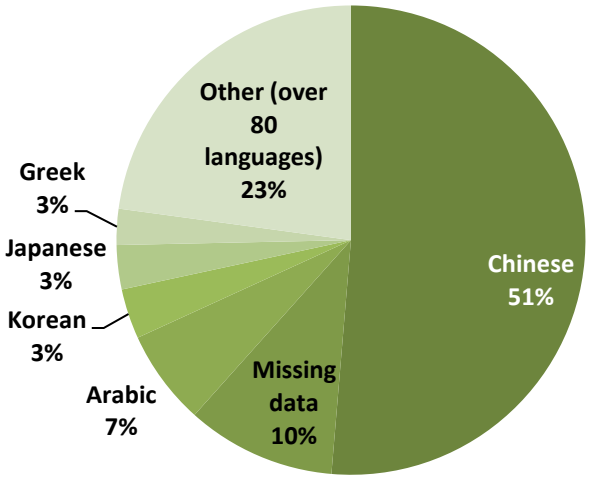

Figure 2 Test takers by first language 2010-2011

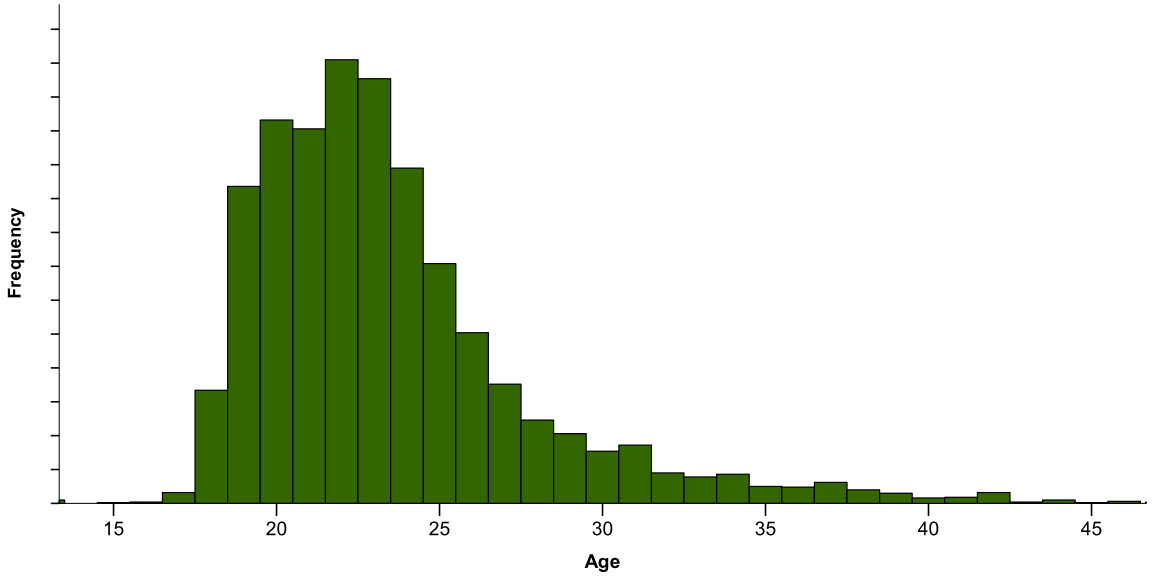

Figure 3 Test takers by age 2010-2011

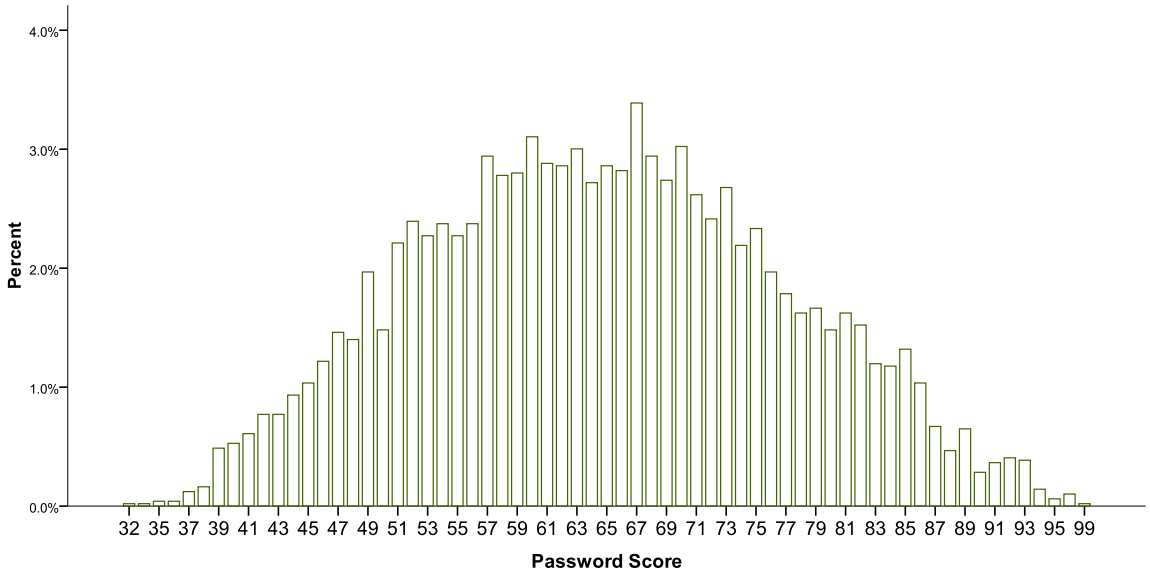

Figure 4 Distribution of total *Password* scores 2010-2011

Descriptive Statistics

Overall

Mean:	59.870
Standard deviation:	16.196
Alpha:	0.916
SEM	1.361

By language

Language	Mean	Standard deviation
Arabic	55.65	13.45
Chinese	57.55	11.97
Greek	68.61	12.55
Japanese	65.95	12.56
Korean	60.83	12.30

Percentile ranks

Percentile	*Password* score
99	90
95	83
90	79
85	76
80	73
75	71
72	69
69	68
66	67
64	66
61	65
58	64
55	63

Percentile	*Password* score
53	62
50	61
48	60
45	59
43	58
40	57
37	56
34	55
31	54
29	53
27	52
24	51
22	50

Percentile	*Password* score
20	49
18	48
16	47
14	46
13	45
12	44
11	43
10	42
9	41
8	40
7	39
6	37
5	34

Reliability

Reliability is an important issue to consider in interpreting and using test scores. The more reliable the scores are, the more confidence we can have that the scores measure test takers' abilities in a consistent manner.

Based on data from a representative sample of over 5,000 ***Password*** tests administered in the period 2010-2011, a statistical estimate of the reliability of ***Password*** (Cronbach's coefficient alpha) is 0.916. A widely accepted rule of thumb for interpreting Cronbach's coefficient alpha is that greater than 0.9 is excellent; 0.8 is good and 0.7 is acceptable (George & Mallery 2003).

This gives a standard error of measurement (SEM) of 1.361. The SEM is an indication of the precision of test scores and signifies how close a test taker's observed test score might be to their true ability on the test.

In the case of ***Password*** the SEM of 1.36 means that we can be 95% confident the test taker's true score is within 2.66 points (+/- 1.96 SEMs) of their observed ***Password*** score. This means that when tests taker's actual points score is in the middle of a ***Password*** grade e.g. 60 points in the 57 to 62 point - inclusive - ***Password*** 5.0 grade band we can be 95% confident that the test taker's ***Password*** result is correct. As the test taker's actual score moves closer to either the upper or lower ***Password*** grade points boundaries there is an increased probability that their true score is ***Password*** 0.5 higher or lower than that reported. Users can be over 99.999% confident that a test taker's true ***Password*** score is not more than 0.5 higher or lower than that reported.

The evidence is that ***Password*** test results exhibit excellent reliability.

Related Documents

These can be downloaded from www.englishlanguagetesting.co.uk.

Studies undertaken by institutions using ***Password,*** show further evidence of the ***Password*** test's reliability and accuracy of results. See the NCUK paper entitled: '***Password*** Predictive Study'.

For more information about the alignment of ***Password*** scores to the CEFR see the paper entitled: '***Password*** and the CEFR'.

The ***Password*** website (www.englishlanguagetesting.co.uk) provides further information and opportunities to view sample ***Password*** tests.