

Relating Password to the Common European Framework of Reference

Summary Report

Centre for Research in English Language Learning and Assessment

University of Bedfordshire

The current suggested correspondence between Password and the Common European Framework of Reference (CEFR) levels is as follows:

A2 Password 3.0 and above

B1 Password 4.0 and above

B2 Password 5.5 and above

This suggestion is based on a study carried out by the Centre for English Language Learning and Assessment at the University of Bedfordshire.

In considering the relationship between Password and the CEFR, it should be kept in mind that Password is a test that focuses on knowledge of the English language – on grammar and vocabulary – while the Common European Framework of Reference (CEFR) embraces language in use in communication. Knowledge of the specific grammar and vocabulary required for communication in any given language lies beyond its purview.

This paper briefly outlines the methods used to relate Password to the CEFR Performance Level Descriptions and the results obtained.

Methods

In carrying out the project, we adopted a range of methods for relating Password to the CEFR in order to compare the outcomes and to provide as full a picture as possible of the degree of correspondence. We have three sources of data for establishing the relationship based on judgements made by a panel of language education experts participating in a one-day standard setting workshop held at the University of Bedfordshire and following recommendations made by the Council of Europe in the publication *Relating examinations to the Common European Framework of Reference: A Manual* (Council of Europe 2009), henceforth referred to as 'the Manual'.

These methods include

- Using the CEFR to rate writing samples collected from Password test takers for whom Password scores are available. These writing samples are not part of the Password test, but are administered at the same time and are used by institutions to supplement Password scores. The Password scores required for each level are obtained by regressing Password scores on the CEFR ratings.
- Two methods for setting standards based on test items: the *basket method* and a variation of the *bookmark method*. Detailed descriptions of these standard setting methods can be found in the Manual.

Participants

In line with the minimum recommended in the Manual, ten participants took part in the linking workshop. These participants represented a range of those involved in some capacity in writing or editing the test (including one participant from English Language Testing Ltd, one from the University of the Arts, one from the University of Bedfordshire and one from Lancaster University International Study Centre) and others who were users of the test, but who had no responsibility for its design. These included participants from Aberystwyth University, Kaplan International Colleges, University College London, the University of Reading, the University of

Southampton and a second participant from the University of the Arts London. All but one of the participants (L1 Japanese) were L1 speakers of English.

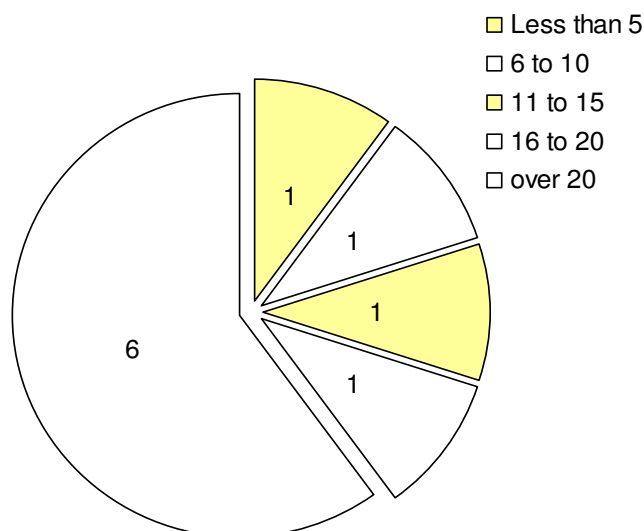


Figure 1 Participant experience of English language teaching in years

ELT Qualifications and knowledge of the CEFR

Eight of the ten had a Masters level degree in TEFL or Applied Linguistics and one of these also held a PhD. Of the two without Masters level qualifications, one held a Postgraduate Certificate in Education in TEFL and the other a Cambridge Diploma in Teaching English Language for Adults (DELTA): eight of the ten held diplomas in TEFL (DELTA, RSA Dip. TEFLA).

All felt they had at least a 'Basic' knowledge of the CEFR: four rating their knowledge as 'Good' and six as 'Basic'. All reported some previous professional experience of using the CEFR, although only one had previously participated in a test linking exercise (for the City and Guilds examination board). Other uses made by participants of the CEFR included materials development and selection; rating scale development, assessment of students and the evaluation of qualifications submitted by prospective students.

Familiarisation

The Manual recommends that whatever their level of knowledge of the CEFR, participants in CEFR linking projects should take part in preliminary exercises to (re)familiarise them with the level interpretations. Although all participants in the Password linking project already had some knowledge of the CEFR, they were asked to carry out a series of exercises to remind them of the Performance Level Descriptions, the scales presented in the CEFR based on 'can-do' statements, and to build shared understanding.

Ahead of the meeting, all participants were asked to review the CEFR, giving particular attention to the Performance Level Descriptions. They were then asked to assign a CEFR level to a series of ten can-do statements taken from the *general linguistic range* (p.110) and *grammatical accuracy* (p.113) scales of the CEFR and to ten grammar and ten vocabulary statements from the DIALANG system (Alderson 2006), which is itself based on the CEFR. These statements have been related to the CEFR levels as part of the DIALANG development.

On the day of the seminar, following procedures recommended in the Manual, the participants were divided into two groups of five and first individually and then in their groups were asked to reconstruct the self-assessment grid on page 26 of the CEFR by reviewing the (randomly ordered) statements in turn and assigning each to a CEFR level. The group members then discussed their decisions and resolved any disagreements. The two groups were then able to compare their decisions against the self-assessment grid and to engage in further discussion in a plenary session.

Results of Familiarisation

Reflecting their knowledge of the CEFR, in responding to the online questionnaire the participants were generally able to assign the 10 CEFR statements to the correct level. Of the 100 judgements made (10 statements assigned by 10 participants), 78 were accurate: the participant correctly identifying the CEFR level of the can-do statement. Of the incorrectly assigned items all but one was placed at a lower level than in the CEFR and all but one was placed into the adjacent CEFR level. For example, four participants placed the statement '*Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people*' at the A1 level rather than at its actual CEFR level of A2. One participant correctly placed all ten statements and three of the ten

statements (two at B2 and one at A1) were correctly placed by all ten participants. The participant who placed the highest number of statements incorrectly (six of the ten) placed five of these at one level below their actual level and the sixth (*Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations*) two levels below. The statement '*Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse*' was placed by five participants at B1 and by five at its actual CEFR level of B2. Otherwise the CEFR level of each statement was correctly identified by the majority of participants. The results are summarised in Figure 2.

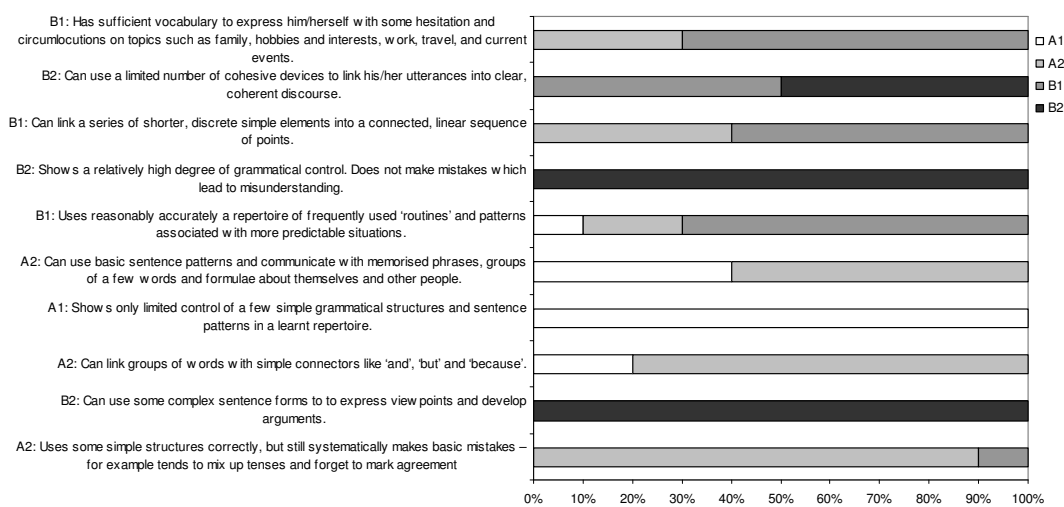


Figure 2 Participant judgements of the CEFR level of 10 can-do statements

The DIALANG can-do statements for grammar and vocabulary did not elicit the same levels of agreement as the CEFR statements. 39% of the vocabulary and 38% of the grammar can-do statements were correctly placed by the participants at the CEFR level identified in the DIALANG project. Again most of the remaining statements were placed at an adjacent CEFR level, but they were somewhat more evenly distributed above and below the intended levels. 40 vocabulary descriptors were placed below their DIALANG CEFR level and 21 above. For grammar, 44 were placed below and 18 above their DIALANG CEFR level. Seven vocabulary and nine grammar items were placed more than one level above or below the CEFR level established for DIALANG.

The familiarisation exercises demonstrated that the participants had a good awareness of the CEFR levels as operationalised by the self assessment grid and

that their judgements of the DIALANG statements were also broadly in line with the previously estimated levels. Where discrepancies arose, the participants in the panel tended to be somewhat harsher than the CEFR in assigning can-do statements to levels.

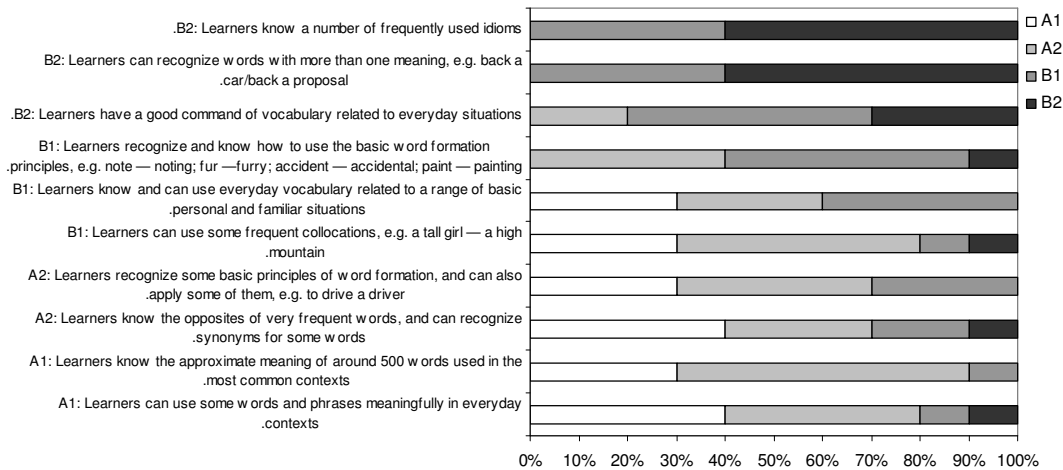


Figure 3 Participant judgements of the CEFR level of 10 can-do statements from DIALANG: vocabulary

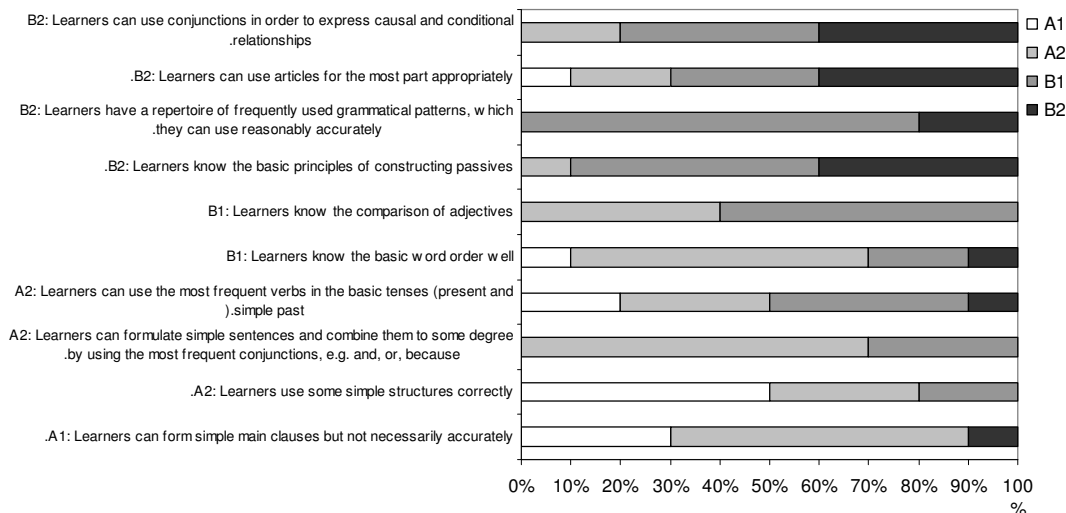


Figure 4 Participant judgements of the CEFR level of 10 can-do statements from DIALANG: grammar
Standardisation

Following familiarisation, the participants, divided into two groups, were presented with one of two sets of ten scripts. Eight of the scripts in each set had been written by Password candidates (in response to tasks administered by certain universities alongside the test) and two were benchmarked samples taken from the CEFRtrain website (www.ceftrain.net) as representative of CEFR levels.

The participants, working independently, used the CEFR scales to assign each script a global CEFR level. Following discussion of results within the groups, the two groups then exchanged scripts and scored the second set. The results for each script for all ten participants were analysed using a multi-faceted Rasch procedure to obtain scores adjusted for rater severity. These scores could then be compared with the Password scores obtained by the same candidates to provide an indication of how Password scores might correspond to CEFR based judgements of writing ability.

A majority of participants assigned each of the benchmark scripts to the correct CEFR level: 58% of the ratings of these scripts were accurate and the remainder were all within one CEFR level of the benchmark rating. None of the raters was identified through the Rasch analysis as misfitting, suggesting that all were consistent in their judgements.

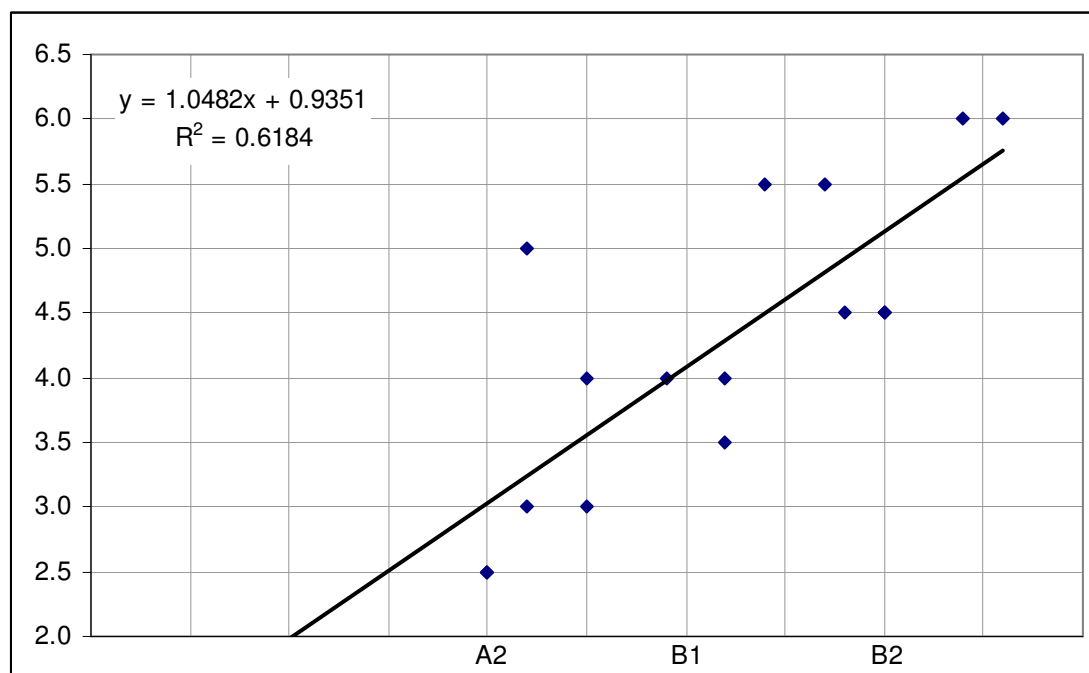


Figure 5 Scatterplot of the relationship between Password scores and CEFR ratings of writing samples from Password test takers (n=16)

As shown in Figure 5 above, there was a close relationship between the Password scores and the participant judgements with a regression coefficient (R^2) of .618. This exercise suggests that a Password score of 3.0 or above is consistent with a CEFR level of A2, Password 4.0 with B1 and Password 5.0 or perhaps, more cautiously, 5.5 with CEFR B2.

The basket method

The basket method has been used in a number of CEFR linking studies including DIALANG. In this approach participants are asked to place each item in a basket corresponding to one of the CEFR levels. So, for example, if an item is put in basket B1, this means that a person at B1 or above should be able to give a correct response to this item.

The method to convert judgments to cut-off scores is based on the reasoning that participants set minimum requirements for each level. In the example given by the Council of Europe, on a 50 item test, if a participating judge places two items in Basket A1, seven in Basket A2 and 12 in Basket B1, then it follows that according to this participant, $2+7+12 = 21$ items should be responded to correctly by any one who is at B1 or higher. This number, the minimum requirement, is interpreted as the cut-off score. In the case of Password, we did not employ a full test form, but adapted the method to a sample of items drawn from the item bank, taking as the cut score the test taker ability corresponding to the average of the item difficulties identified by participants with the relevant level. Each participant thus made a judgement about the level of each of 45 items of known difficulty (calibrated on the basis of the performance of over 1,000 pilot test candidates) drawn from the Password item bank.

Again using multi-faceted Rasch procedures to adjust for rater severity, estimates could be obtained by this means for the ability level of a test taker who might be expected to score at each CEFR level. In this case, one rater and two items were identified as misfitting (with standardised infit mean square greater than 2.0) and were excluded from the analysis.

The participants were moderately successful at ordering the items according to difficulty. There was a correlation (r) of .63 between the participants' ratings and the calibrated difficulty values for the items. Taking the mode of the participant judgements for each item, nine of the 45 were placed at A2, 24 at B1 and 12 at the B2 level. The average difficulty of the items considered to be at each level is taken to indicate the likely level of ability of test takers at the corresponding CEFR level.

The Rasch model used in constructing Password allows us to estimate the probability of a test taker of known ability responding correctly to any given item drawn from the bank. It is thus possible to calculate the score on Password that

corresponds to a greater than 50% chance of success on the selection items identified with each level. Applying this method suggests that the cut score for the A2 level should be Password 3.0, for B1, Password 4.5 and for B2, Password 5.5.

The bookmark method

Following the bookmark method, participants are instructed to start with the lowest standard (here A1), and go through the test items as they are presented in order from easy to hard, deciding for each item whether a test taker at this level would be likely to give a correct response. At some point, as they proceed through items of increasing difficulty, the participant will judge that the borderline test taker is more likely to give an incorrect than a correct response. A bookmark (real or symbolic) is placed next to the item that falls at this point. The participant then switches to the next higher standard (A2 in this case) and continues work from the next item in the sequence.

In this exercise, each participant was presented with an ordered set of items running from the easiest to the most difficult on the basis of data from over 1,000 Password test takers. They were asked to identify the point in this set at which test takers at each CEFR level would pass from being more likely to answer items correctly to being more likely to answer incorrectly (A1, A2, B1 and B2). The most appropriate cut point between levels was taken to be the average of the difficulty of the most difficult item that test takers at a given level should answer correctly and the easiest item at the level above. An average was then taken of all ten participants' judgements to arrive at a third set of recommendations for cut scores.

On this basis, the participants suggested that an A1 level test taker would be able to answer 10% of the items considered in this exercise, an A2 level learner 29%, B1 62% and a B2 test taker 88%.

Using the predictive power of the Rasch model to extrapolate from the items analysed, the results indicate that test takers with Password 3.0 and above should be considered to have a level of language knowledge commensurate with A2 level, CEFR B1 would be consistent with Password 4.0 and B2 with Password 5.5. Again the results would seem to correspond reasonably well with the outcomes from the other two standard setting approaches adopted.

Conclusions

On the basis of the procedures described above, it is suggested that Password scores correspond to levels of performance on the CEFR as follows:

A2 Password 3.0 and above

B1 Password 4.0 and above

B2 Password 5.5 and above

However, it is acknowledged that although results from Password are consistent with CEFR ratings, Password is not comprehensively a test of language performance as operationalised in the framework and so should be taken as a relatively indirect measure of CEFR level.

It should also be understood that any linking project is of necessity tentative. As outlined in the Manual other techniques for standard setting are available and different participants or different methods might yield different estimates of the relationship between test and framework. It can also be argued that all tests and frameworks are social constructs whose interpretation will shift over time and so it is necessary to revisit and, if necessary, revise recommendations as the test or the framework develops or as new techniques for linking become available.

References

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Modern Languages Division, Council of Europe.

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual*. Strasbourg: Language Policy Division, Council of Europe.

Alderson, J.C. (2006) *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London: Continuum.