…AND BACK TO MULTIPLE CHOICE! LARGE-SCALE TESTING OF PROFICIENCY IN ENGLISH: AN EXPERIENCE[1]

Irina Argüelles Álvarez and Iciar Pablo-Lerchundi, Universidad Politécnica de Madrid
Email: irina@euitt.upm.es

Abstract: The aim of this paper is to demonstrate the validity of a specific multiple choice test to situate a student above or below a B2 proficiency level according to the Common European Framework of Reference for Languages (CEFRL). The research process and the results, not the product, are the focus of this work. After carrying out a pilot study of the test with 214 students, results regarding its reliability and validity are statistically analyzed and explained in detail. It is demonstrated that a properly designed multiple choice test can discriminate whether or not a student has reached the required B2 level according to the CEFRL.
Keywords: testing, test construction, test validity, test reliability, test preparation.

Título en español: …¡Y de vuelta a la opción múltiple! La evaluación de la competencia en lengua inglesa a gran escala: una experiencia

Resumen: El objetivo de este trabajo es demostrar la validez de un test del tipo "opción múltiple" para situar a un estudiante por encima o por debajo de un nivel de competencia B2 según el Marco Común Europeo de Referencia para las Lenguas (MCERL). El proceso y los resultados más que el producto final son el centro de este artículo. Tras un estudio piloto con 214 estudiantes, los resultados relativos a su validez y fiabilidad se analizan y se explican en detalle. Se concluye que una prueba de elección múltiple correctamente diseñada puede discriminar si un estudiante ha alcanzado o no el nivel B2 según el MCERL.
Palabras clave: evaluación, diseño de pruebas, validez de pruebas, fiabilidad de pruebas, preparación de pruebas.

INTRODUCTION

In September 2009, the new four-year degrees in the area of Telecommunications adapted to Bologna began in the Technical School of Telecommunications Engineering at the Universidad Politécnica de Madrid, UPM (Technical University of Madrid). Like the rest of the cases in the UPM, these new degrees substitute the three or five-year engineering degree programs which have long been the ones adopted in Spain. The UPM has made

[1] **Date of reception**: 23 April 2012
Date of acceptance: 30 August 2012

an enormous effort to renew the institution in order to adapt its organization and also its educational model to current times, as proved by a number of actions carried out over the last several years[2].

With regard to the students' education for their incorporation in a global professional world, the UPM has included a university-wide compulsory subject in the new curriculum aimed at preparing students for international academic and professional situations. The specific subject which will be the responsibility of the Department of Linguistics Applied to Science and Technology has been called "English for Professional and Academic Communication". In most cases, it will be offered during the last semester of each of the different engineering degree programs. The important detail here and the starting point of the research presented in this paper is that the University requires the students to certify a B2 level as described by the Common European Framework of Reference for Languages (CEFRL) in order for them to have the right to enroll in that compulsory subject which, in the case of our center, the Technical School of Telecommunications Engineering, is programmed for the seventh semester of the new four-year degree.

The specific fact which triggered and justifies the subsequent study presented herein is that a pilot experiment carried out in September 2009 (Argüelles *et al*. 2010a) confirmed the teachers' general intuition that most of the students at the UPM have not reached that B2 proficiency level at the time they start their university studies. To be more specific, the results from this pilot study showed that only 20.85% of the 250 students participating in the pilot placement interviews and tests had reached a B2 proficiency level as described in the CEFRL. With these percentages in mind, an important number of considerations and questions became the focus of our research, which has been carried out during the past year. Some of these considerations and questions were clearly related to teaching and learning issues that needed to be solved from both the perspective of educational curriculum development and a more organizational point of view, where, regretfully, important contextual constraints had to be taken into account. Some others were related to the placement process itself, its validity and reliability, as it had been carried out in the pilot study, and the future need for a large-scale placement system effective in practical terms of time, human, technical and other necessary resources.

As a response to the first group of considerations concerning immediate teaching and learning needs, different actions were straightforwardly initiated in October 2009, which are still in progress. These actions are aimed at filling the gap between the proficiency level the students are required to have and the proficiency level each of the students actually demonstrates. The institution at its different levels immediately started on-line general English programs for e-learning (UPM), developed face-to-face courses of general English for levels below B2 (Department of Linguistics Applied to Science and Technology) and offered varied alternatives in order for the students to keep in contact with the language throughout their studies (Centers and Innovation Groups). Among the alternatives offered to students by the Centers through their Innovation Groups to bridge the gap between

[2] As the change affects every aspect of this large and old institution, it is difficult to summarize here all the notorious improvements that have taken place at the University over the last few years. However, that is not the focus of this paper. To obtain more information about the University, its evolution and its offer, visit http://www.upm.es

their starting proficiency level and the required B2, the Integrated Language Learning Lab (ILLLab) project was begun in the Technical School of Telecommunications Engineering. The ILLLab initiated two main blocks of activities to integrate the English language into the routine of the students' engineering studies. The first block is a cultural program including a cinema forum, "English through Films," together with a series of conferences by native speakers, "Around the World in English." The second block is content and language-integrated activities to be included in different compulsory b-learning or e-learning subjects of the four degrees (Argüelles *et al*. 2010b).

Returning to the second group of considerations related to the students' placement process and the need to establish a reliable and valid large-scale proficiency exam, steps have also been taken in that direction. It is in fact the development process of such a test that is the focus of this paper. The process rather than the product is analyzed hereafter as it is worthwhile examining how the surrounding conditions, the specific context where the test is undertaken and its aims influence the decisions to be made. In what follows, first, the decision-making process involved in preparing the test is described; later in the discussion section, the alternatives chosen are justified; and finally, the strengths and weaknesses of the final product are explained in view of the results.

TEST SPECIFICATIONS: FORMAT AND GENERAL TEST DESIGN

Our specific context, the number of students who must certify a B2 level to enroll in the subject "English for Professional and Academic Communication" and the heterogeneous background of the more than sixty teachers in the Department of Linguistics lead us to opt for an automatic correction test made up of multiple choice- type questions. If it is well designed, the test can include items assessing various aspects of the foreign language, from the most traditional in this type of test, grammar and vocabulary, to more functional or even pragmatic aspects. Here are some examples of items extracted from the test to illustrate how the different language aspects are assessed as well as to clarify what we mean by "functional" or "pragmatic" aspects:

- Grammar

Mozart was born in Salzburg _____ 1756.
a) in b) on c) at d) in the

- Vocabulary

It can take you half an hour to get to the train station in the _____ hour.
a) rush b) busy c) hectic d) crowded

- Functional/pragmatic aspects

(On the phone)
A: Good morning. Is Mr. Smithson in, please?
B: One moment, please, I'll _____.
a) connect you b) put you through c) let you in d) take your message

Although for the same practical reasons of application, the test does not include a listening comprehension section or an interview, previous studies (Argüelles *et al*. 2010a) lead us to establish an initial hypothesis which presumes that most Spanish students at

university who demonstrate a high level of proficiency in these aspects of language would obtain similar proficiency results in direct listening and speaking tests. The limitations of these types of tests are therefore understood and assumed for practical reasons of application, and the extent to which they affect the students' results will be presented and analyzed in the results and discussion sections below.

The test is designed for the access of students to a compulsory subject, "English for Academic and Professional Communication." The University requires the students to have a B2 level of proficiency according to the CEFRL in order to enroll in this subject and successfully follow a course in academic and professional communication. As the students are not supposed to have taken a previous course of this type, items addressed to test academic or professional English are discarded.

The test is designed to check that the students have the required minimum previous knowledge and not to place them within a scale. In other words, the test must indicate whether the student has reached the required B2 level and, therefore, the sort of test to be developed to meet that requirement is a proficiency test, not a placement one. The students do not receive feedback or information about their level of proficiency within the levels established by the CEFRL. However, a message is sent to students regarding whether or not they have reached the threshold of the level required to enroll in the subject "English for Academic and Professional Communication." Here, we suggest developing another free access and voluntary placement test for students, with feedback and detailed information about the levels to help them have a clearer idea of their level of proficiency inside the scale. This knowledge will eventually enable the students to plan their English learning during their university studies.

To achieve the expected result and in order for the multiple choice test to asses vocabulary, structural aspects, use of language and reading comprehension, the specific text of the activities and the options available need to be adapted to the aims, and there needs to be a sufficient number of items that test each of these aspects. The layout of the test shows two differentiated parts although the test is not explicitly divided into those two parts. The first part that evaluates aspects of grammar, consists of 65 individual items followed by the four options a, b, c and d. The second part of the test focuses on aspects of language usage, vocabulary and reading comprehension and comprises three texts of 10 to 15 points each, for a total of 35 points. The four options are presented in a gap in the position where the word or the words are missing: a, b c or d. For security purposes, a resource repository of alternative test items that assess the same aspects of language is utilized in order to present the same standard of difficulty in the test's different versions.

The test is designed for the b-learning platform Moodle independently of the possibility that, in the case of lack of technological support, the different versions are printed and delivered as a paper and pencil exam to the students. The items are stored in the platform, organized in blocks by language category (grammar, vocabulary, use of language and reading) and by topic within these categories, so that for each of the versions the program will choose a predetermined number of activities from each of the blocks. That is to say, the program will make a semi-random selection of items, limited by the condition that it chooses one item from each of the groups.

The four-option multiple choice test items are adapted from a corpus of texts and tasks selected from general English course books which have been correlated to the CEFRL, covering B2 and extending towards a C1 level. From the corpus, the core vocabulary, grammatical structures and functions, and the difficulty of the texts for the level are established. In a first revision, once the activities have been adapted, native-speaker teachers of English with experience in testing check that the items are clear and unambiguous. At the same time, they verify that only one of the options provided can in fact be correct. Apart from the correct option, the other three options are adapted to the following scheme: one answer seems very likely although it cannot be possible and the other two are not possible. Whenever possible, one of these last two options represents common error tendencies typical of Spanish learners of English as a foreign language. These tendencies have been noted during years of teaching experience and tend to be a result of interference from the native language, as well as other factors.

Eg. 1: I always leave home early _____ avoid the morning rush hour.
a) in order to b) so that c) so as to d) for to
Eg. 2: I know it isn't lunch time _____, but I'm starving!
a) yet b) already c) still d) no longer

The pilot test consists of 100 items selected from approximately 1,000 validated items. In this case, for research reasons, the 100 items selected are the same for all the students taking the pilot test. For future exams, the selection of items is to be made randomly by the program for each of the students taking the exam. The test has been tried out in an experimental situation with first-year students in the UPM School of Telecommunications. These students are representative of the sort of students who will take the test in the future. Then, results are studied to reach conclusions concerning the test reliability and validity in the context where it is presented.

RESULTS

A total of 240 incoming students at the School of Telecommunications took the test. The test is administered and supervised by teachers in computing rooms set up with approximately 30 computers each. Two groups of students complete the test: one in the morning and another in the afternoon. Of the 240 students taking the test, 36 did not finish it. In some cases this was due to the students' low level of proficiency, in other cases it was due to more technical reasons: the student did not close the questionnaire after finishing it or the student closed the questionnaire before finishing it. Therefore, a total of 214 tests are taken into consideration for the statistical analysis.

In what follows, a summary of the results of the test concerning its reliability and validity is presented.

Internal Reliability

The reliability of a test is the extent to which the results can be considered stable or consistent. A reliable test is one in which a subject should get a similar score if he was able

to repeat it. Hence, reliability refers to the quality of the test as a constant, and therefore reliable, measuring tool (Brown 1988: 98, 99).

Split-half Method

The reliability of a test can be estimated by calculating its reliability coefficient. One of the most common methods to test reliability and obtain its coefficient is the split-half method. Here, the correlation between two halves of a test (usually the even-numbered items and the odd-numbered ones) is calculated. With this method, the Spearman-Brown coefficient or the Cronbach alpha is obtained. Both Spearman-Brown and Cronbach alpha coefficients range from 0.0 to 1.0, where 0 means that the test is not reliable. These numbers provide the user with the percentage of the variation in the true scores which is related to a natural variation in scores, and shows the reliability or stability of the test. The remaining percentage that cannot be attributed to the variation of the true scores is the error or unsystematic variation.

The Statistical Package for Social Sciences (SPSS) software is used for this analysis and the results show a Spearman-Brown coefficient of 0.87 and a Cronbach alpha of 0.89. This means that 87% to 89% of the variation of the test is due to variation in the true scores, so the test can be considered consistent. Only an estimated 13% to 11% of the variation cannot be accounted for.

Correlation between the parts

The test is laid out in two well differentiated parts: the first part focuses on grammar aspects while the second is centered on aspects of use of language, vocabulary and reading. As such, it is important to analyze to what extent these two parts show similar results. If the test is well designed, the correlation between the parts should be high**.** This would mean that a student who gets a high score in the grammar part also gets a high score in the second part of the test, whereas a student with low scores in the first part also gets low scores in the second part. The Pearson correlation coefficient shows the extent to which two sets of scores covary (vary together) (Woods *et al*. 1986).

The correlation coefficient ranges from -1.0 to +1.0, where -1.0 means that the relationship between the two sets of scores is exactly opposite. That is, if a subject gets a high score on one set, he gets a low score on the other set and vice versa. Secondly, 0.0 means that there is no correlation between the parts. Finally, a correlation coefficient of +1.0 shows a direct relationship where high scores in one part are related to high scores in the other part. Therefore, a high positive correlation coefficient would mean that a student who gets high scores in the grammar part also gets high scores in the second part of the test, which demonstrates an appropriate design of the test as a whole.

The Pearson correlation coefficient between the scores obtained in part one and those in part two of the test is analyzed by means of SPSS and the result is 0.81, which is statistically significant (see Table 1). This means 81% concordance between the two parts of the test.

Table 1. Pearson Correlation between part one and part two of the English Proficiency Test.

	Part two (use of language, vocabulary and reading)
Part one (grammar)	0.813**

**p< 0.01

Validity

The validity of a test is the extent to which the test is actually measuring what it claims to measure. Traditionally, three main types of validity have been considered once the reliability of the test has been established. First, content validity refers to the extent to which the items are a representative sample of the contents that the test claims to measure. Second, construct validity is related to the idea that a test tries to measure something that is not observable, for example, proficiency in English. Therefore, construct validity usually requires an experiment to prove that the test actually measures the non-observable construct that it claims to measure. Third, criterion-related validity implies an external criterion. This criterion is established and accepted as a reliable measure of the same construct that the designed test is measuring (Brown 1988: 102-105). In previous sections, the reliability of the test was analyzed to see whether it provides stable and constant scores. In what follows, the validity of the test is analyzed.

The criterion chosen to test the validity of the test was an interview with an examiner, always the same judge to avoid bias derived from inter-rater reliability. The examiner, based on the CEFRL B2 level descriptors, establishes whether each of these students has the required B2 level. This method had already been proved reliable for establishing a level of proficiency in English in Argüelles *et al*. (2010a) as the results showed that the level assigned to a student from an interview was statistically identical to the level of proficiency obtained by the same student in the vocabulary and grammar part of the Oxford Placement Test (Allan 2004).

As for establishing validity in our test, a statistically representative sample from the 214 students who completed the multiple choice test was selected for the interview. The examiner, unaware of the previous results of the 31 students, held a personal ten-minute interview with each of these students. The correlation between the results in the interview and the test was analyzed by means of SPSS and a validity coefficient of 0.83 was obtained. This correlation is statistically significant.

Table 2. Pearson Correlation between Oral interview and English Proficiency Test.

	English Proficiency Test
Oral Interview	0.825**

**p< 0.01

As the interview is a direct test which evaluates the students' speaking and listening skills, the results were later used to reach conclusions regarding the extent to which the results from the multiple choice test can be used to measure other more communicative

abilities. This correlation will also be used to justify the validity of the multiple choice test and to establish the cut-off score that distinguishes B2 from non-B2 students.

Cut-off score

From the experiment, the results show that 29% of the 31 students interviewed reached the B2 level according to the CEFRL, whereas 71% did not reach this level of proficiency. Derived from these results, our concern now is to set the cut-off score in the multiple choice test, which will allow the institution to make decisions, as the students who score over the given cut-off score will be considered *masters* (students who have reached the B2 level) or *non-masters* (students who have not reached the level).

Taking into account the results obtained from the oral interview and correlating them with the scores in the multiple choice proficiency test, the cut-off score could be established in the range from 6.9 to 7.1. If we take into account the global number of students who took the multiple choice test, a cut-off score of 6.5 to 6.6 could be established, as 29% of the students obtained a score of 65/66 or higher.

However, according to Bachman (2001: 75), it is also possible to establish a cut-off score for making decisions on the basis of the distribution of scores from a norm-referenced test. For example, to enter a program in the case that we wanted to be highly selective, the cut-off score could be set at two standard deviations above the mean. This in our case means that the cut-off score could be set at 8.8, which is much higher than the preliminary results from the experiment.

Whenever a mastery / non-mastery decision is made, two possible types of errors can occur (Bachman 2001: 75). In our case, a false positive classification error would occur if we classified a student with a level lower than B2 (a high B1 level) as a B2 level student. Therefore, setting a cut-off score of 6.5 to 7.1 could result in false positive classification errors. On the other hand, if we establish the highly selective cut-off score of 8.8, a false negative classification error could occur, such that students with an English level of B2 are classified as having lower levels.

As the established cut-off scores are too divergent and in order to avoid these two classification errors, an intermediate cut-off score could be a possible solution. If the mean is established between the cut-off score based on the correlation between the oral interview and the multiple choice test and the resulting highly selective cut-off score, a score between 7.7 and 8.0 is decided as the final cut-off score.

DISCUSSION

It is understood here that this proposal could be highly unpopular. Given that assessment and evaluation in contexts of higher education tend to introduce direct techniques and task-based and other more formative approaches, there are some important facts to clarify at this point.

First, the proposal is made for a specific context where a B2 level must be proved by students enrolling in a course of professional and academic English. Here it is not expected to obtain learning results, to place a student within a scale or to diagnose possible gaps in

their previous learning but to evaluate the students' proficiency in English in order to meet an administrative requisite to permit their admission in a specific subject.

Second, multiple choice tests are a good alternative for increasing reliability, and the design and development of the test were conceived to maintain good levels of criterion-related validity. Although the validity of this kind of test to measure a level of proficiency is usually considered uncertain, according to the results presented in the above sections, the test shows robust statistical results concerning its concurrent and predictive validity. As content and construct validity are more conceptual than statistical (Davies *et al*. 1999), the domain and the theoretical model presented was aimed at giving a precise response to the situation and the needs that frame our specific context.

Third, it is not intended here to suggest or demonstrate that an indirect test of this type can in any case substitute direct tests of different skills, but rather to present it as a practical tool where other alternatives are difficult or impossible to carry out. The aim of the test**,** which is in fact to certify that the student has reached the required B2 level**,** justifies the assumption of the costs associated with the possible classification errors. Estimations for minimizing both types of classification errors (false positive and false negative) have also been carried out and explained in the above sections.

To summarize, an examination of the statistical results presented in the body of this paper shows high levels of reliability and validity of this test to measure what it claims to measure in order to situate a student above or below the given B2 level of proficiency. During the investigation, two aspects of the testing process were central to this piece of research: the first was the establishment of a cut-off score. The second, which was derived from the first, was the extent to which an eventual erroneous classification of students above or below their level of proficiency might affect the program and the evolution of the students within it.

CONCLUSIONS

It can be concluded, from a preliminary analysis of the data,that the test designed to verify the students' level of proficiency is working correctly. At this point, it is important to emphasize that this test has some limitations stemming from the type of evaluation that is being carried out. Although the contextual circumstances led us to conclude that this type of test is the most reliable in our case, the face validity of the test with regard to what it aims to evaluate is in fact affected. It must be remembered that a B2 level, like the rest of the levels as described in the CEFRL, is specified on the basis of a series of competencies that the speaker of the language must demonstrate, whereas a test of grammar and vocabulary is not the ideal solution for evaluating competencies. The decisions made with regard to the use of a multiple choice test were justified previously.

We must highlight that the information provided by the test addresses the students' knowledge of the grammar and the vocabulary associated with this B2 level according to the CEFRL. The test also obtains information regarding a minimum degree of reading comprehension and some other sub-skills that students should have developed by this level. Based on the initial pilot study, it is presumed that students who demonstrate this level of proficiency would eventually pass a competency type test of the same level as the one

assessed here in a high percentage of the cases. Nevertheless, this test does not assess a B2 level of competence directly according to the CEFRL. To improve the test validity, we recommend including an interview which would be carried out by teachers trained for that specific aim, a limited board that would assume only that function. This recommendation is difficult to attain in the short term due to the time needed to train teachers and the large number of students to be assessed at the University. Indeed, this should represent the greatest effort and organizational challenge for the coming years.

ACKNOWLEDGEMENTS

We want to acknowledge the Universidad Politécnica de Madrid for its support of this research project funded by its annual Innovation Project Program and Professor Ceo-DiFrancesco (Ph.D., Xavier University, Cincinnati) for her revision and encouraging comments to the final draft of this paper.

REFERENCES

Allan, D. 2004. *Oxford Placement Test*. Oxford, Oxford University Press.

Argüelles I., E. Martín, R. Herradón, G. Balabasquer, C. Ortiz. 2010a. "Análisis y descripción de la variación en el nivel de competencia en lengua inglesa de los estudiantes que ingresan en los nuevos grados de ingeniería de telecomunicación" *AESLA*, Vigo.

Argüelles, I., J. Sendra, M. Millán, R. Herington, J. Blanco, R. Herradón.2010b "TIC y aprendizaje integrado de contenidos técnicos y lengua inglesa en la EUIT de Telecomunicación" *Interdisciplinariedad, Lenguas y TIC: Investigación y Enseñanza*, Valencia.

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford, Oxford University Press.

Brown, J. D. 1988. *Understanding research in second language learning*. Cambridge, Cambridge University Press.

Common European Framework en http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp

Davies, A., A. Brown, C. Elder, K. Hill, T. Lumley and T. McNamara.1999. *Dictionary of language testing*. Cambridge, Cambridge University Press.

Grupo de Innovación Educativa Multidisciplinar (2009) *Integrated Language Learning Lab* http://illlab.euitt.upm.es

Statistical Package for the Social Sciences (SPSS) http://www.spss.com/?source=homepage&hpzone=nav_bar

Woods, A., P. Fletcher and A. Hughes. *Statistics in language studies*. Cambridge, Cambridge University Press.