UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

http://revistas.um.es/ijes

Item and test construct definition for the new Spanish Baccalaureate final Evaluation: A proposal

JESÚS GARCÍA LABORDA*
Universidad de Alcalá
ELENA MARTÍN-MONJE
Universidad Nacional a Distancia (UNED)

Received: 21 September 2013 / Accepted: 22 October 2013

ABSTRACT
The current English section of the University Entrance Examination (PAU) has kept the same format for twenty years. The Bologna process has provided new reasons to vary its current format, since the majority of international reputed tests usually include oral sections with both listening and speaking tasks. Recently the Universidad de Alcalá (Madrid) was granted the funds to suggest and experiment the new format through the OPENPAU project. This paper justifies the introduction of the Spanish Baccalaureate Final Evaluation in terms of a re-design of the tasks in line with the Communicative Approach and the Common European Framework of Reference and also taking into account the distinctive features of a language test in order to be valid and useful (construct validity, reliability, impact, practicality and quality management). Furthermore, it provides sound suggestions for a computer-based delivery which could be done through the use of mobile devices. The article concludes with the belief that the new diploma may serve to overcome the deficiencies of the current exam if rigorous studies are undertaken and research-based decisions are reached.

KEYWORDS: testing, University Entrance Examination, construct, computer-based exam

RESUMEN
La actual sección de inglés dentro de la Prueba de Acceso a la Universidad (PAU) ha mantenido el mismo formato durante veinte años. La reforma universitaria de Bolonia ha proporcionado nuevas razones para modificar el actual formato, ya que la mayoría de los exámenes internacionales de reconocido prestigio incluyen secciones orales con tareas de comprensión y expresión. Recientemente la Universidad de Alcalá (Madrid) recibió fondos para proponer y experimentar con un nuevo formato a través del proyecto OPENPAU. Este artículo justifica la introducción de la Evaluación Final de Bachillerato en lengua inglesa en términos de un rediseño de tareas en línea con el Enfoque Comunicativo y el Marco Común Europeo de Referencia, que también tome en cuenta las características distintivas de un test de lenguas de manera que sea válido y útil (validez del constructo, fiabilidad, impacto, practicidad y gestión de la calidad). Además, proporciona sólidas sugerencias para la puesta en marcha de un examen informatizado que podría realizarse a través de dispositivos móviles. El artículo finaliza con el convencimiento de que esta nueva evaluación puede servir para superar las deficiencias de la prueba actual si se llevan a cabo estudios rigurosos y las decisiones alcanzadas se apoyan en la investigación realizada.

PALABRAS CLAVE: examen, PAU, constructo, examen informatizado

Address for correspondence: Jesús García Laborda. Dpto. de Filología Moderna, Facultad de Filosofía y Letras, c/ Trinidad, 3, 28801, Alcalá de Henares-Madrid, Spain. Tel: 34 918855041 E-mail: jesus.garcialaborda@uah.es

IJES, vol. 13 (2), 2013, pp. 69-88
Print ISSN: 1578-7044; Online ISSN: 1989-6131

I. INTRODUCTION

Even though foreign languages have been a requirement to enter the Spanish public universities for over 20 years (Fernández Álvarez & Sanz Sainz, 2005), most regional boards (“Consejerías de Educación”) have followed almost exactly the same format in the English section of the University Entrance Examination (PAU): a 200 word text with comprehension questions, a few “fill in the gaps” grammar exercises and a writing essay (usually 120 - 200 words) (Fernández Álvarez & Sanz Sainz, 2005), and it was not until the turn of the century that the Catalonian educational board of education decided to include multiple choice listening tasks. This test design, based on Oller’s cognitive theory on Unified Competence (Oller, 1979, 1983), which is based on the theory that competence is a unified set of interacting abilities and thus test takers evidence similar levels of competence across all the tested skills, is internationally rejected today because students do not perform evenly across the different language skills and testing components (García Laborda, 2010). However, despite the relevant work by McNamara (2000) on the need of discrete point testing, its perpetuation is probably due to the low cost of assessing objective tasks, especially the written ones, and the continuity of an educational system that is neither open nor brave enough to admit structural changes (García Laborda & Fernández Álvarez, 2010).

When language tests are used in different contexts and different subjects it is hard to predict their validity and fairness (Kane, 2010). It is true that the global demands brought by the Bologna Process (1999) involve greater mobility for workers and students, and thus the need to assess their language competence especially speaking mostly by standardized tests that can be used to obtain adequate inferences and to improve the teaching of foreign languages in high school. Furthermore, the education authorities seem to advocate for well-established examples of foreign language testing (such as TOEFL, IELTS, etc.) (Kohkhan, 2012; Stoynoff, 2009; Zahedi & Shamsaee, 2012) or previously designed models (Kim & Craig, 2012) to meet that need, when considering changes in the current format. However, it is also evident that not all the test adaptations may be valid and a more in-depth analysis is required. The new test also needs to meet the adequate standards, as will be seen in the following sections of this paper. Not only that, the implementation of listening and speaking tasks in this criterion-referenced high stake test calls for an adequate calibration, in order to avoid negative implementation side effects.

In line with other similar tests, there are three main issues that items and tasks need to address and that will both have an impact in the classroom (washback) and shape the construct: Objectivity, fairness and directness. In other words, the Baccalaureate Final Evaluation (BFE henceforth) that will foreseeably substitute the current PAU should have a similar degree of difficulty and provide similar inferences across socio-cultural and geographically different groups. Additionally, like other tests, it is shaped by a number of factors, not least of all the cost of assessment. Bachman (1990) states that it is hard to consider what “objectivity” may mean in language testing, since there are not “pure objective or subjective” tests, and also because no skill is absolutely applicable to either characteristic. Thus, the traditional thought that speaking and free writing alone are subjective enough to challenge their use in “objective tests” may be wrong.

Since the impact of the current exam is so important in the decision of the candidates’ admission to certain university degrees (such as Medicine and others), teachers fear that introducing these tasks in a new BFE may have a negative effect on their students’ scores (García Laborda & Fernández Álvarez, 2012). Additionally, speaking has been the most neglected skill in foreign language learning in Spain (Talaván, 2010). Thus, although teachers would probably support the implementation of speaking tasks, they also fear them (Díez-Bedmar, 2011; Fernández Álvarez & García Laborda, 2012; Martín-Monje, 2012). As a consequence, it is necessary to present the new construct adequately to the different stakeholders in order to avoid the natural reluctance to adopt a new test. Accordingly, the final goal of this paper is to introduce and contextualize the tasks in the test construct for an adequate implementation and delivery of the BFE.

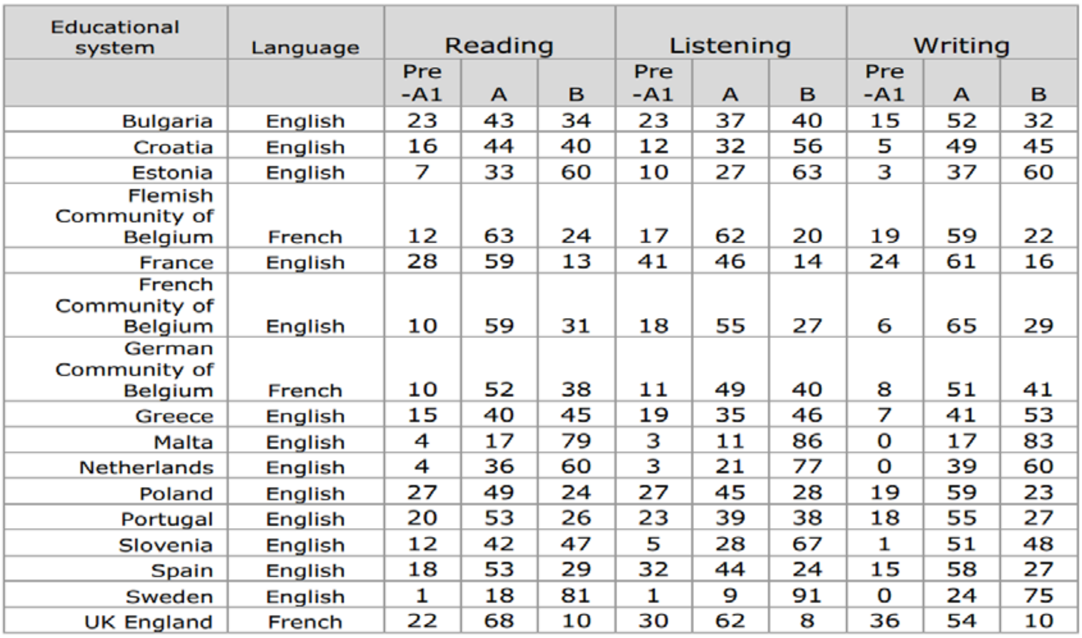
In this sense, this paper differs from others which tackle this issue from an experimental perspective. This piece of research takes a more descriptive approach (similar to Taylor & Gernanpayeh, 2012), since its main goal is to suggest significant changes for the future BFE.

2. JUSTIFICATION OF THE INTRODUCTION OF THE ENGLISH SECTION IN THE SPANISH BACCALAUREATE FINAL EVALUATION

2.1. English as a foreign language in Spain

Weir (2013) states that language tests change to adapt themselves to language teaching methods, modern linguistics, theories of second language acquisition, but institutional tests also need to accommodate to the different stakeholders’ needs. This is especially true when deficiencies can be observed in the educational system or teaching outcomes such as the national second language competence across international studies. In the last years, there has been an increasing interest in changing the PAU test due to a number of reasons. There are two main types of justifications to do so: the current language competence nationwide and the need to improve the general features of the test. In relation to the first, Spanish students underperform in English when compared to other European countries especially in the English as a Foreign Language which is currently the most determinant Lingua Franca in the World. For example, in 2011 only 19% of Spanish people between 25-64 years old believed that they had a proficient knowledge of English (Eurostat Press Release, 2013) and many believed that the language education they had in high school had a very limited effect in their lives (García Laborda, Bejarano, & Simons, 2012). This language deficiency was also observed in the European Survey on Language Competence (ESLC, European Commission,

2012) and has a direct impact in the current concerns of the Spanish Ministry of Education, Culture & Sports (MECD). The ESLC collected information about the foreign language proficiency of students in the last year of lower secondary education or the second year of upper secondary education, depending on the country. The data related to Spain refer to the first option, students in the last year of lower secondary education, which coincides with the end of their compulsory studies. A summary of the conclusions reached can be seen in Table 1 below.



Educational system	Language	Reading			Listening			Writing		
		Pre -A1	A	B	Pre -A1	A	B	Pre -A1	A	B
Bulgaria	English	23	43	34	23	37	40	15	52	32
Croatia	English	16	44	40	12	32	56	5	49	45
Estonia	English	7	33	60	10	27	63	3	37	60
Flemish Community of Belgium	French	12	63	24	17	62	20	19	59	22
France	English	28	59	13	41	46	14	24	61	16
French Community of Belgium	English	10	59	31	18	55	27	6	65	29
German Community of Belgium	French	10	52	38	11	49	40	8	51	41
Greece	English	15	40	45	19	35	46	7	41	53
Malta	English	4	17	79	3	11	86	0	17	83
Netherlands	English	4	36	60	3	21	77	0	39	60
Poland	English	27	49	24	27	45	28	19	59	23
Portugal	English	20	53	26	23	39	38	18	55	27
Slovenia	English	12	42	47	5	28	67	1	51	48
Spain	English	18	53	29	32	44	24	15	58	27
Sweden	English	1	18	81	1	9	91	0	24	75
UK England	French	22	68	10	30	62	8	36	54	10

Table 1. First foreign language – Percentage of students achieving levels at the end of their compulsory education by skills and educational system

It is clear that Spain is in a disadvantaged position against most of the other European countries. Looking at the data skill by skill, the results in the only oral skill assessed, listening, are even more worrying, since they show that one third of our students (32%) are below the most basic Common European Framework of Reference level, which is A1, beginner.

In relation to speaking, 12th grade students (year of high school graduation), evidence even more serious deficiencies. In this sense, the National Institute of Educational Evaluation (INEE) studied in 2012 the speaking competence of 975 students in seven regions in Spain (Asturias, Aragón, Islas Baleares, Castilla La-Mancha, Comunidad Valenciana, La Rioja and Madrid). The individually delivered test included two types of tasks: a set of basic questions (informal examiner-student conversation) to elicit simple personal responses and an oral description of a picture followed by one or two questions related to the visual clue. Table 2 below shows that results obtained according to the rating criteria: topicality, grammar accuracy, fluency, interaction and speech coherence.

	Rating criteria					
	Topicality	**Grammar Accuracy**	**Fluency**	**Interaction**	**Speech coherence**	**TOTAL**
1st part	65.08	60.39	64.66	68.17	68.51	63.65
2nd part	61.06	55.53	62.81	64.49	65.66	60.80
Overall Pass	60.22	54.77	61.39	63.82	64.57	61.39

Table 2. Percentage of Spanish students at 12th grade who obtain a pass grade at B1 level in the CEFR in Speaking

As can be seen, just over half of the students have a CEFR B1 level (Independent user, intermediate) when finishing high school. These results could be considered acceptable if it were not because at this point of schooling students have taken between 9 and 12 years in the foreign language. Additionally, it is debatable whether the B1 competence level has a real potential in terms of employability and use for pursuing studies in higher education. The ideal scenario would be a cohort of students who have a B2 level at least. (Independent user, upper intermediate), in order to cope comfortably with the demands of a society in which student and worker mobility across countries is on the rise. In fact, most schools abroad require a B2 to follow higher education courses or a B1 to register in their intensive language programs, and a First Certificate in English (CEFR B2) is the minimal certification that many companies demand in Spain. As a consequence, if this situation is to be improved, tests could have the potential to trigger the necessary change.

2.2. Research principles

This paper aims to define and describe the tasks to be included in the construct of the English section of the BFE of general English ability and the delivery of the test as understood by the research team. The paper also intends to show how these tasks can contribute to the general improvement of the English competence in Spain. In so doing, the paper relies mostly on an interactional perspective (Van Compernolle, 2011). The paper also considers the framework for developing and validating language tests by Weir (2005), and considers the latest developments in computer-based language tests through both desktops and mobile technology.

As opposed to other approaches that tend to describe the construct first, the following sections will begin by justifying the mere existence of such construct based on the expected impact in reference to the current needs of the Spanish educational system that have been described previously. Therefore, the methodological approach is based on analysing the needs in order to suggest ways to address them. Due to this, the following literature review mainly focuses on the current reality in Spain, which certainly can be extended and applied to other countries.

3. TEST VALIDITY IN THE CONSTRUCT OF THE SPANISH BACCALAUREATE FINAL EVALUATION

According to Messick (1989: 13), validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". The process of test validation has become essential in the field of Language Assessment in the past decade (Lim, 2013)

In reference to the test validity and its features, one of the most significant ones is the impact produced by tests in education. From a theoretical perspective, Messick (1989, 1996), and Alderson & Wall (1993) suggested that exams have a powerful effect on how teachers teach their classes. The effects can also be observed in the learners' attitudes, in the teachers' methodological approach and activity selection, the classroom setting and context, the introduction of grammar or communicative contents and the range of materials to facilitate language learning. The impact of exams on all of these aspects and, in general terms, on how classes are instructed was called "washback" or "backwash" (it is beyond this paper to discuss the differences and henceforth we will address it as "washback"). There are clear signs that washback has existed in Spain for many years and that Bachillerato teachers (covering the last two years of high school) consider that it is extremely important to teach towards the PAU (García Laborda & Fernández Álvarez, 2012), even if they feel that students are missing learning possibilities (Amengual, 2009, 2010; García Laborda & Fernández Álvarez, 2012). However, well directed washback is a potentially relevant factor to improve the role and quality of language learning in Spain (Amengual, 2010). Besides, it can also have a significant importance in the individual's learning, education and professional career (through access to higher education or simply by providing him with communication skills to succeed in his/her job). Thus, as a global phenomenon, washback can have a social impact through providing citizens with the language knowledge to work in and out their own country (e.g. bilingual nurses are currently demanded in some places in the European Union). This is part of what Bachman and Palmer (1996) called "test impact" and certainly is based on what is going on in the classroom but its effects clearly go far beyond.

The new Spanish educational reform LOMCE (Ley Orgánica para la Mejora de la Calidad Educativa), which is intended to be passed before the end of 2013 (http://www.congreso.es/public_oficiales/L10/CONG/BOCG/A/BOCG-10-A-48-4.PDF), introduces an educational system in which regular assessments and two high stakes assessments (after 10th and 12th grades) serve to take decisions about the students. Of these, the latter, BFE, will provide valuable information on the student's language competence.

Furthermore, the test also serves as a control test for the access to the public universities as mentioned above.

In this context is it crucial to pinpoint all the aspects that need to be changed and improved in the current English section of the PAU, so as to make it a more meaningful and valid test that fits into the forthcoming BFE. In order to do so, it is necessary to look at the four distinctive features of a language test first mentioned by Bachman in a set of unpublished seminars in Cambridge at the beginning of the 1990's: construct validity, reliability, impact and practicality (VRIP) (Lim, 2013). Most of these aspects have already been identified in previous publications as needs of the Spanish Educational System. Table 3 shows a summary of the literature review that covers VRIP in Spain.

Features	Needs	References
Quality Management	• Studies are needed in all the aspects of the PAU • Management should be run professionally because high stakes decisions are taken	Díez-Bedmar, 2011b García Laborda, 2010
Practicality	• The test is very practical and easy to administer but should fulfill the good test features • The rating process should also be checked to assure that rating is fair • Computers should be used to • To facilitate the delivery process • To achieve a more efficient and rapid rating of certain section • To enrich the task input for productive tasks • Other delivery devices could be used for the speaking tasks in PAU like mobile phones • Teachers require training to achieve higher teaching performance	Fernández Álvarez & García Laborda, 2012 García Laborda, Giménez López, & Magal Royo, 2011 García Laborda & Magal Royo, 2007 García Laborda & Magal Royo, 2009 García Laborda, Magal Royo, & Enríquez Carrasco, 2010 Martín-Monje, 2012a
Impact	The test has to improve its current negative washback	Amengual, 2009, 2010 Díez-Bedmar, 2011a
Reliability	• The test should be enlarged because the current size does not provide evidence enough • to take evidence-based decisions • The option between a criterion or norm based test needs to be revised • Serious piloting prior to delivery would be desirable • Alignment with other external tests should be assured	García Laborda, Bakieva, González Such, & Sevilla Pavón, 2010
Construct validity	New tasks should be included that are more adequate to current communicative teaching methods such as speaking and listening	García Laborda, Bakieva, González Such, & Sevilla Pavón, 2010

Table 3. Literature review related to PAU test VRIP

Test validation is an ongoing activity; it has to be repeated over time in order to check that the processes are being followed. It is because of this that Lim (2013) has added a quality management dimension to the VRIP scheme (which has been accounted for by the OPENPAU project as well), so that it captures the "how" of managing the necessary processes of test validation (Figure 1).

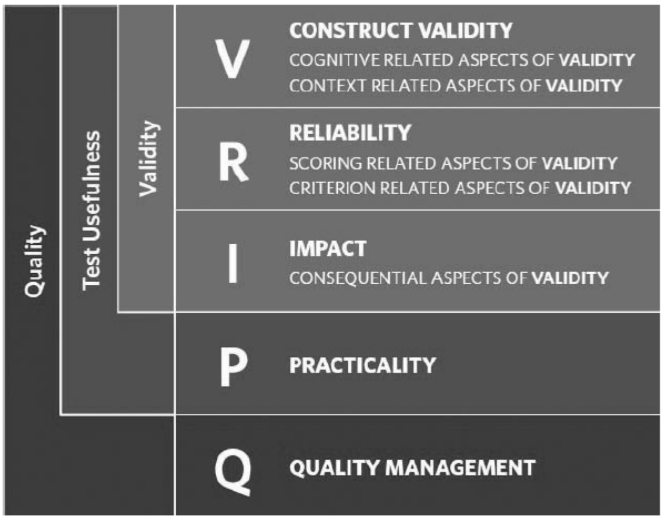

Quality
Test Usefulness
Validity
V
CONSTRUCT VALIDITY
COGNITIVE RELATED ASPECTS OF VALIDITY
CONTEXT RELATED ASPECTS OF VALIDITY
R
RELIABILITY
SCORING RELATED ASPECTS OF VALIDITY
CRITERION RELATED ASPECTS OF VALIDITY
I
IMPACT
CONSEQUENTIAL ASPECTS OF VALIDITY
P
PRACTICALITY
Q
QUALITY MANAGEMENT

Figure 1. VRIPQ framework (Lim, 2013)

Apart from these inherent needs based on the qualities of a test, other realities justify the changes at a national level. The increasingly generalised use of ICT (Information and Communication Technology) in educational contexts calls for a different approach to test administration and delivery. In the last years computer testing has achieved a higher status (Brown, 2012; Chapelle & Douglas, 2006) in test delivery for a number of reasons of which economy, speed of delivery and rating, administration and delivery facility and other aspects make it a versatile tool for high stakes assessments (García Laborda, 2007). Computer language testing has acquired its unique interest especially in the new century, although previous works by Brown (1997) were done in the field of adaptive computer based testing (Chapelle, 2001; Stahl, Bergstrom, & Gershon, 2000; Weinberg, 2001; Xiao-Fan, 2000). Alderson (2000) explained the importance of use of CD Roms in language testing and a few years later, Chapelle, Enright, & Jamieson (2008) provided valuable insights into the IBT TOEFL^{TM} online test. However, computer based language testing acquired full maturity after the introduction of the Computer based TOEFL, and especially after the integration of the IB TOEFL in 2005. By then, the Universidad Politécnica de Valencia was working on a new computer based testing tool called Herramienta Informática de Evaluación Online (HIEO), a pioneer tool which has been followed by a number of projects in Spain. In relation with this, García Laborda (2010) has repeatedly suggested that the integration of computers in the PAU would benefit all the test stakeholders. The following section deals with the possibility of implementing a computer-based exam and suggests models for the different tasks that would enable an automated correction in most cases.

Furthermore, another significant change has been suggested in the last years to improve the information and inferences of the PAU test. One of the most significant criticisms of the PAU has been the limited information that it provided on specific realities associated to the school districts and, more importantly, individual high schools. By improving the test through the increment and adequacy of the test construct and locally based delivery information it would be more adequate and reflect better on the specifics of each high school. In line with this, changing the PAU for the BFE has the following advantages:

1. The results justify remedial and immediate decision and actions in the teaching. These actions can also be taken faster.
2. It permits to establish qualitative and quantitative differences among students. For example, in the current PAU no analysis beyond a basic rating is provided.
3. It would favor the integration of cognitive approaches to testing along with socio-constructivist approaches to formal assessment delivery (Norris, Leighton, & Phillips, 2004; Poehner, 2007).
4. Based on point 3, this type of test facilitates the distinction of the different aspects of the evaluation such as process, delivery difficulties, impact on the high school as a whole and especially its language courses and teaching methods as well as the test results.
5. It can diagnose the different improvements in instruction across time.
6. It can consider the school’s assessment and test process and their implications in the overall competence improvement.
7. It improves the relationship between formal and informal language learning.

4. CHOOSING THE RIGHT TASKS

Once the test validity and its features have been looked into, the next step in this research has been to consider the construct definition and test tasks. The first decision to be taken was to opt for using either discrete-point or integrated tasks. So far, most of the regional Spanish PAU models have tended to use isolated tasks which only demand one skill at one time, but modern testing, especially computer based, has shown the importance of combining tasks. In his previous project, PLEVALEX, García Laborda (2006) preferred the discrete-point system while the more recent PAULEX (Martínez Saez *et al.*, 2011) favoured the integration of skills. The current OPENPAU project is working towards the integration of skills through a computer based test approach as well as a paper-based one.

Since our current research has led to this dual approach, integrating test skills at a low cost demands serious decisions such as the kind of prompts to be used or whether a direct or indirect approach will be used to assess speaking tasks. According to Carr (2011) direct tests require the immediate use of the skill assessed, while Sauvignon (2001) mentions that indirect tasks tend to be used for productive skills through related tasks, since they measure the ability or knowledge that underlies the skill that is being assessed. However, Bachman (1990) dismisses such a differentiation, since it seems awkward for him to make a distinction between them. As for the Spanish context, the OPENPAU project acknowledges that the traditional PAU test is based on a direct measurement of writing and also has a clear purpose, to measure the candidate's ability to perform in reading at an intermediate level (García Laborda, 2010) (the PAU has not been openly and explicitly linked to the CEFR and therefore makes no reference to its standardised levels). However, it fails to reliably measure any other skills, such as spoken comprehension or production. It is mainly due to this fact that the validity of the foreign language section is challenged. As a consequence, the OPENPAU project has chosen to use integrative semi-direct tasks in which the students record their performance. These are also considered the most adequate activities for the aim of this project, since they allow for flexibility in rating and delivery, which is something that direct tasks lack, especially the listening and speaking ones.

4.1. Qualities of usefulness

The English section in the PAU requires reliability, an assurance that the test will measure what it is intended to do. Nevertheless, one of the problems of the current format is that it has no descriptors and candidates are not currently informed of the specifications tested. The only information teachers and students are provided with is an exam model, together with the marking criteria. The latter only specify the marks allocated to each exercise, the skill measured in that specific part of the test and how potential errors will be penalised. This is partly due to the fact that PAU administrators are usually not professionals in language testing. In fact, the PAU coordinator in foreign languages does not need to have any previous experience or knowledge in testing since this position is awarded to any university professor without an objective appointment or assessment of previous work in language testing. As a consequence, the test lacks almost any of the VRIP features previously mentioned to make it useful. In fact, the exam has long been criticized for such lack of control and trialing. It is quite common to find instances of mistakes in the PAU exam in the press (Vida, 2013), and the foreign language exam is no exception in this regard.

It is still not clear whether the desired changes and improvements will be implemented in the foreseeable future, but a serious revision of the test is required and there are several options to keep its cost as low as it is at the moment. Nevertheless, the OPENPAU project itself establishes that the test requires an in-depth study of its validity. In that sense, it is necessary to come up with an outline of the features of a good test. With that aim, the researchers have followed the qualities of usefulness proposed by Bachman and Palmer (1996) as presented in table 1.

Quality	Measure
Reliability	If the test is computer based, automatic rating will be preferred. When human intervention is necessary, objective items will be preferred. Consistency of scoring becomes the ultimate goal in this case. Current practices do not respond to the desirable homogeneity of rating / scoring decisions.
Authenticity	The test needs to resemble, present and demand the kind of target language use that candidates will require in real-life situations. Today, the test is a combination of the most traditional grammar exercises with unconnected tests and unlinked compositions.
Construct Validity	It is necessary to provide the test with a high degree of appropriateness that allows making inferences on the candidate’s competence based on the score. Since the section today lacks a strong construct definition, the interpretation is meaningless. That implies that the scores can hardly be transferred to any benchmarking threshold (like the one included in the European Framework of Reference for the languages) and thus the interpretation of the scores may not relate to any precise stage of language acquisition.
Impact	The impact needs to be positive. Current practices in Spain have led to examples of language teaching in the last two years of high school in which learning the language is neglected and all the efforts are put into developing the test skills needed for the PAU.
Practicality	The test is practical as delivered today but since the previous four qualities are not met, its overall practicality can be easily questioned.

Table 4. OPENPAU measures for quality of usefulness (based on Bachman & Palmer)

4.2. Task format

For the purposes of this paper only the tasks that the authors intend to use or at least trial have been described. In this sense, the progressive stages in the test description such as the plan and design of the test itself, the specifications for the tasks, the test writing process and piloting and other issues such as validation, test administration and so forth have been omitted.

When considering tasks and the methodological rationale behind them, the Communicative Approach has been adopted (Richards & Rodgers, 2001) taking into account the latest theoretical developments in foreign language teaching that have been implemented after the publication of the CEFR, with the development of the Action-oriented Approach (Council of Europe, 2001). One of the key innovations of this framework is the switch from the traditional four skills (listening, speaking, reading and writing) to eight language activities, categorized into four types: reception (listening and reading comprehension), production (oral and written), interaction (oral and written) and mediation (oral interpretation and written translation).

Besides, a worthwhile side effect of the implementation of the CEFR in the item and test construct would be the possibility of providing a CEFR level for each of the students taking the university entrance exam. This would be a definite strength of the new test: at present students can only achieve those qualifications taking examinations with institutions that are outside the official Spanish education system (e.g. TOEFL from the English Testing

Service or First Certificate from the University of Cambridge) and these external exams come as an extra cost for the candidates. Thus, being able to attain those qualifications by simply taking the new BFE would be a major improvement. Not only that, if the new test is well constructed, students will be able to show a "whole language competence" by only testing some aspects, as it has been suggested in literature (Newman, 1985; Oller, 1979). It is because of all these reasons that a special effort has been made in the construction and selection of tasks for the new exam, which are put forward in this paper. Former research projects have attempted to use the teachers' ideas to implement the new tasks (García Laborda & Gimeno Sanz, 2008) but after a moderating process this proved to be a far too traditional approach, with an excessive amount of open questions to be possibly considered for a computer-based exam. Table 5 shows the proposal put forward back then:

Skill	Task	Automated correction
Reading	1. Reading and check True/False response followed by a justification 2. Short communicative opinion based answers 3. Multiple choice questions	Yes No Yes
Writing	1. A 130-150 word composition based on personal attitudes and opinions	No
Listening	1. Watching a mini-clip followed by open questions 2. Watching a mini-clip followed multiple choice questions	No Yes
Speaking	1. Watching a mini-clip and answer short questions 2. Giving a 2 minute mini-talk with the aid of the visual prompt provided	No No

Table 5. Former proposal for computer-assisted PAU exam (based on García Laborda *et al*., 2010)

With this model less than half of the tasks would enable an automated correction, which is not desirable, although it is a known fact that a reliable language test will never permit a 100% automated correction. However, there is still room for improvement. The items for the receptive language activities could be designed in such a way that they all make automated correction possible, even if the productive language activities must remain open. This taxonomy has been taken as the starting point of the present study, which is planning to develop an improved item construct which takes into consideration the guidelines of the CEFR. Table 6 shows an overview of the items included, which will be explained in more detail in the following sections.

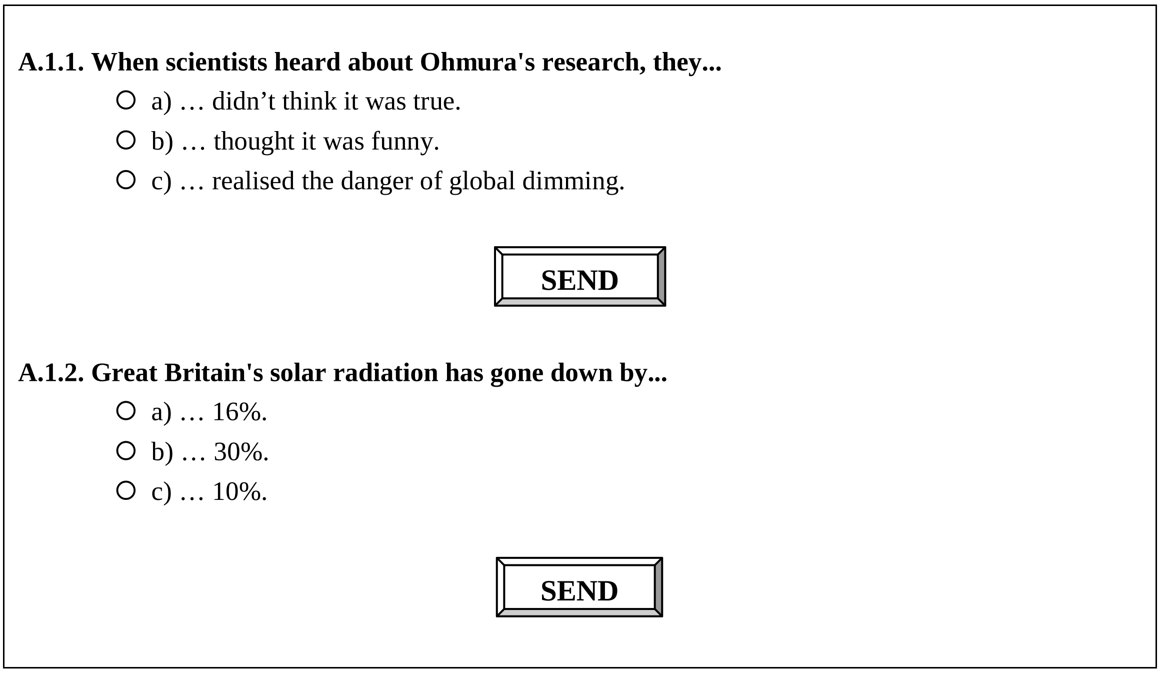
Language activity (CEFR)		Task	Automated correction
Reception	A. Listening comprehension	1. Multiple choice questions 2. Fill in the gaps activity	Yes Yes
	B. Reading comprehension	1. Reading and choosing True/False response followed by a justification 2. Multiple choice questions	Yes Yes
Production	C. Oral production	1. A 2 minute mini-talk using a given prompt	No
	D. Written production	1. A 130-150 word composition based on personal attitudes and opinions	No

Interaction	E. Oral interaction	1. Watching a mini-clip and answering short questions posed by the examiner	No
	Written interaction	N/A	N/A
Mediation	Oral interpretation	N/A	N/A
	Written translation	N/A	N/A

Table 6. New proposal for computer-assisted PAU exam with CEFR language activities

4.2.1 Reception tasks

The reception tasks comprise listening and reading comprehension and have been constructed so as to speed up the correction process. These items enable a computer-based exam or even paper-based ones which use answer sheets to be scanned and marked through machines. The listening comprehension has two activities: multiple choice questions and a fill in the gaps part which should be carefully devised in order to avoid ambiguity, permitting exclusively one-word answers. The reading comprehension is taken from the model shown in table 6, choosing only the activities that are apt for automated correction. Figures 2 and 3 show examples of proposed listening and reading comprehension activities.



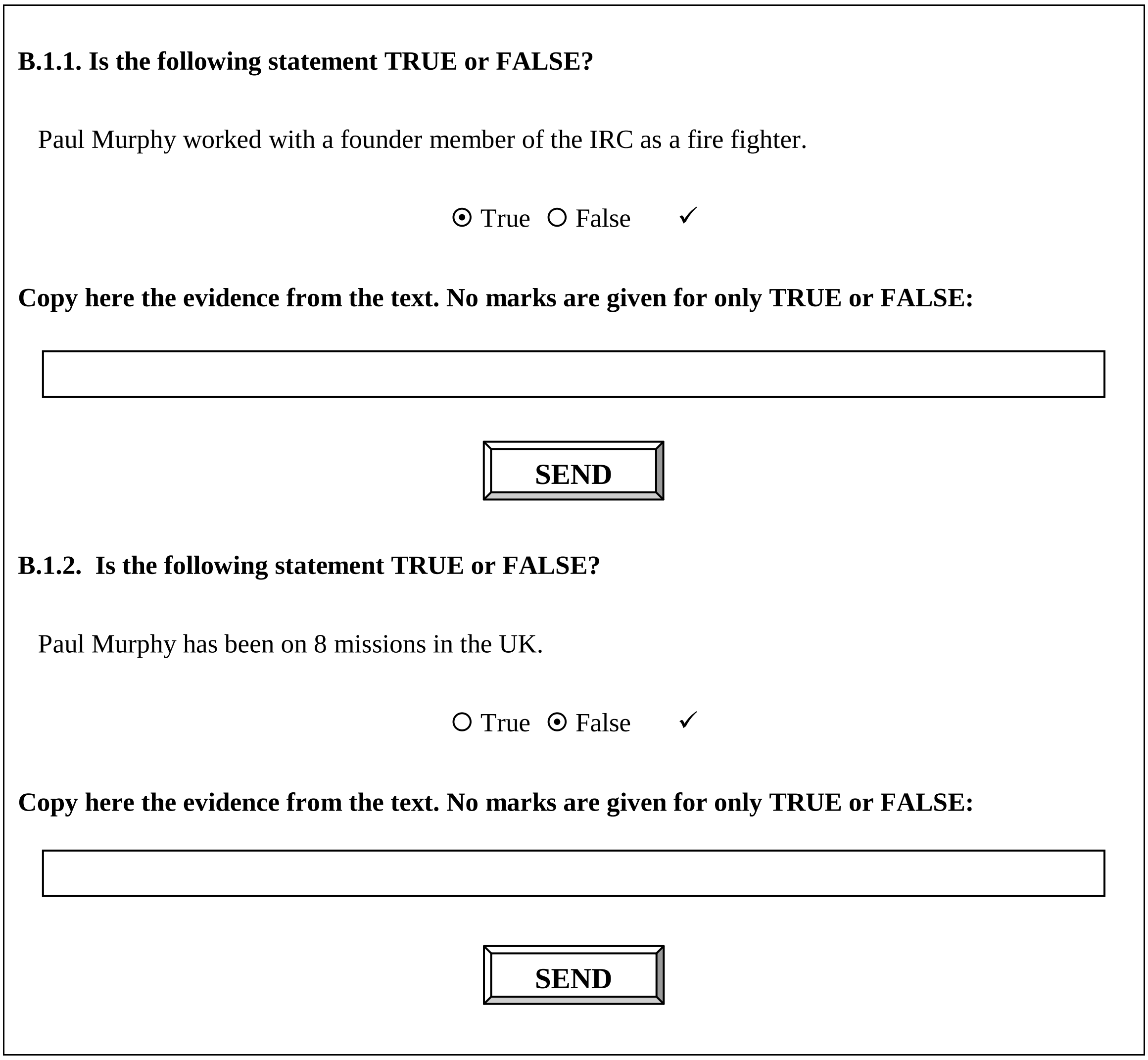
A.1.1. When scientists heard about Ohmura's research, they...

○ a) … didn’t think it was true.
○ b) … thought it was funny.
○ c) … realised the danger of global dimming.

A.1.2. Great Britain's solar radiation has gone down by...

○ a) … 16%.
○ b) … 30%.
○ c) … 10%.

Figure 2. Example of computer-based multiple choice activity in the listening comprehension section



B.1.1. Is the following statement TRUE or FALSE?

Paul Murphy worked with a founder member of the IRC as a fire fighter.

● True ○ False ✓

Copy here the evidence from the text. No marks are given for only TRUE or FALSE:

B.1.2. Is the following statement TRUE or FALSE?

Paul Murphy has been on 8 missions in the UK.

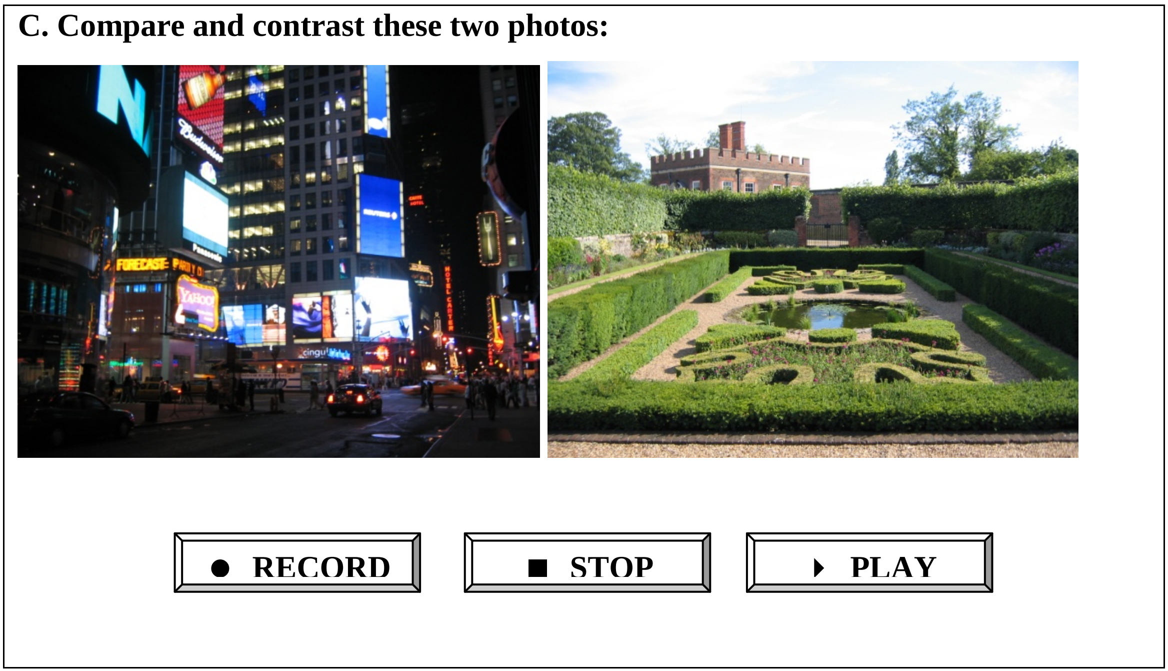
○ True ● False ✓

Copy here the evidence from the text. No marks are given for only TRUE or FALSE:

Figure 3. Example of computer-based True/False activity in the reading comprehension section

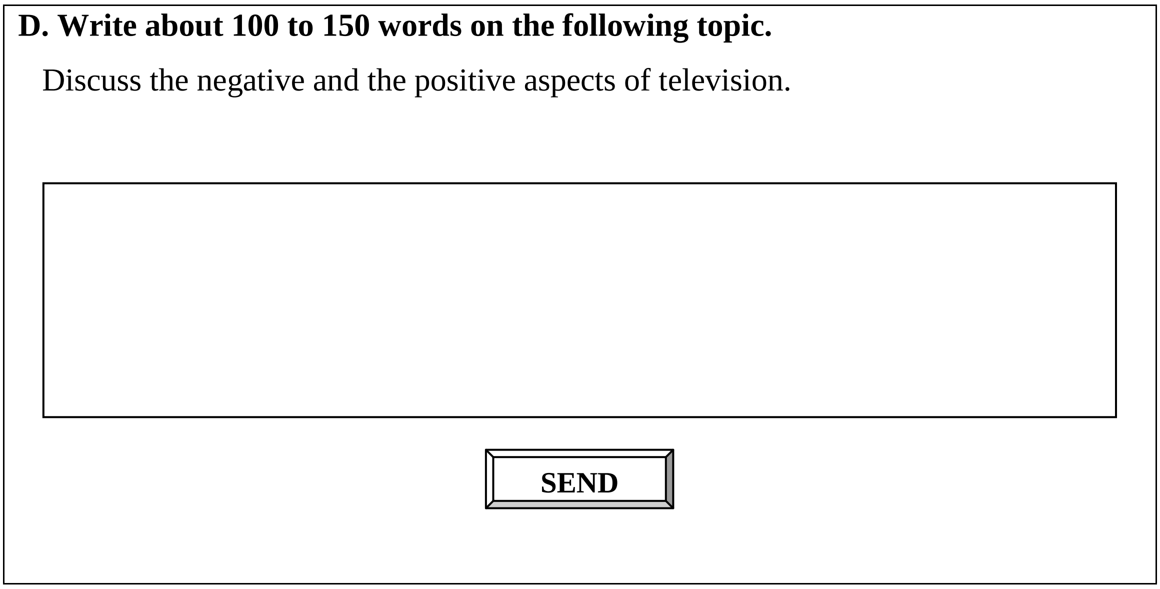
4.2.2. Production tasks

Oral production activities such as the proposed in this project will probably be the hardest and most costly to implement in the new exam. However, no language test can be considered complete without evaluating this competence. The written production activity, on the contrary, does not differ much from previous models, with a 130-150 word composition, which should include letters and creative writing. Figures 4 and 5 show examples of written production and oral production.



C. Compare and contrast these two photos:

Figure 4. Example of computer-based oral production



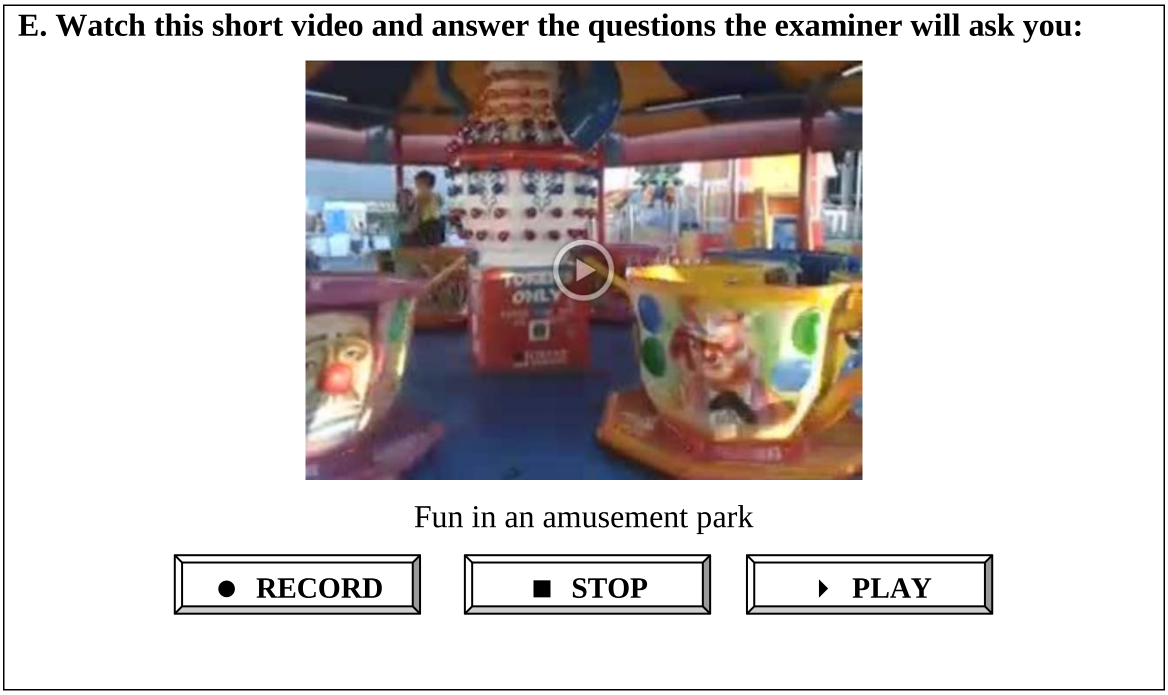
D. Write about 100 to 150 words on the following topic.

Discuss the negative and the positive aspects of television.

Figure 5. Example of computer-based written production

4.2.3. Interaction tasks

The same comments regarding oral production can be said about oral interaction. Our suggestion is a short dialogue between examiner and candidate, prompted by the viewing of a mini-clip. Written interaction is really not feasible, since according to the CFER it includes chatting online, exchanging notes, correspondence by letter, e-mail, etc. and it is not viable to do that in an exam situation.



E. Watch this short video and answer the questions the examiner will ask you:

Fun in an amusement park

Figure 6. Example of computer-based oral interaction

4.2.4. Mediation tasks

Mediation tasks entail oral mediation –simultaneous or consecutive interpretation- and written mediation –translation, summarizing in the second language, etc.- and have not usually been considered part of a foreign language exam. Furthermore, the communicative approach encourages the use of second language only in the teaching and learning of foreign languages.

5. ALTERNATİVE COMPUTER DELİVERY IMPLEMENTATION THROUGH MOBILE DEVICES

Mobile learning, understood as "learning mediated via handheld devices and potentially available anytime, anywhere" (Kukulska-Hulme & Shield, 2008: 273) has been on the rise in the past decade, especially in the modality of blended learning –combining face-to-face and online learning- (Martín-Monje, 2012b). The forthcoming changes in the BFE make it an ideal moment to include this technological development and implement MALL (Mobile Assisted Language Learning) into the test. CALL (Computer-Assisted Language Learning) as such seems to be passé and the 21st century demands the integration of cutting-edge technology –such as smartphones, iPads, etc.- into language learning. Specially, two aspects of mobile devices have been appealing for educational contexts from the start: on the one hand, the convenience of their portability and on the other hand, their widespread use among the population, especially the younger generations (Godwin-Jones, 2011). In countries such as the US there are even government policies that encourage students to bring personally owned mobile devices to their schools in what has been called the BYOD (bring your own device) initiative (US. Department of Education, 2010). Such a scheme should be encouraged to be deployed in Spain. In fact, there has been some successful piloting, such as the PAULEX project, which adapted the PAU exam to mobile devices (García Laborda & Giménez López, 2010; Giménez López, García Laborda & Magal Royo, 2011) and shows that a mobile-assisted English exam would be feasible.

6. CONCLUSIONS

A deep change of the foreign language section in PAU is justified for two main reasons: the poor results of the Spanish students on national and international surveys, and the psychometric features of the test. Many authors have shown the weaknesses of the current model that go from the current lack of research to the test construct itself. This article addressed the most relevant ones. The inclusion of speaking tasks is a must, but the extension of the current tasks to improve the test validity or the revision of the test reliability are also necessary. Among all the aspects, usefulness has become the cornerstone of the change mostly because the test currently provides few inferences of success in both the professional and in higher education. Since high stakes decisions are obtained from the evidence collected from the PAU, it is necessary to develop a stronger and fairer construct that can also inform all the stakeholders of the evolution and changes in language learning. As it is, the current test has limited validity. Changes should also affect how the test is delivered. In this sense, the in-school delivered a Baccalaureate Final Evaluation is more meaningful and can also serve to inform better of the individual and school realities and thus permit a better analysis to implement changes in the contents and teaching procedures in the school, and thus facilitate a positive washback. However, the new diploma should only be the last step in a better designed longitudinal testing plan in which assessments are not the “devil advocates” but a tool to improve the current competence in foreign language.

In relation to test construct definition, tasks build on previous research in the field and take a step forward, linking the Spanish university entrance exam to the CFER and creating a complete, sound, balanced test that measures up to similar exams across Europe. The OPENPAU has suggested many of these changes and is in an advanced stage. It has also given sound suggestions for computer based delivery, which would also reduce the cost of delivery and favor faster achievement reports, and is currently experimenting alternative approaches such as the used of iPads and tablet PCs for the test. Naturally, there will be some trialing and feedback from all the stakeholders to be collected, in case some amendments and/or adaptations are needed.

To conclude, testing is called to have a significant potential in the improvement of languages nationally. If rigorous studies and well taken decisions are reached, the new diploma may serve to overcome the deficiencies of the current situation. In this sense, the

OPENPAU project provides with some justified suggestions but what is truly needed is a real institutional interest and, more than else, a deep understanding of the positive effects of testing in education. In line with this, clearer explanations to both students and teachers would definitely be an asset in the new and promising educational system.

ACKNOWLEDGEMENTS

The researchers would like to express their gratitude to the Ministry of Research and Innovation of Spain (MICINN) for supporting the development and implementation OPENPAU research project (FFI2011-22442) with cofounding with FEDER funds under the 2008-2011 plan. The authors would also like to thank the Spanish Ministry of Education for the funding of part of this through the Proyecto Estudios y Análisis in its call of 2011 to the project «El sistema de acceso a la universidad: Propuestas en la gestión, decisión e inferencias en el área de lenguas extranjeras». We would also like to thank the anonymous reviewers who helped us to improve this paper.

REFERENCES

Alderson, C. J. (2000). Technology in testing: The present and the future. *System*, *28*(4), 593-603.

Alderson, J.c. & Wall, D. (1993). Does Washback Exist? *Applied Linguistics*, *14*(2), 115-129.

Amengual, M. (2009). Does the English Test in the Spanish University Entrance Examination influence the teaching of English? *English Studies*, *90*(5), 582-598.

Amengual, M. (2010). Exploring the Washback Effects of a High-stakes English Test on the teaching of English in Spanish Upper Secondary Schools. *Revista Alicantina de Estudios Ingleses*, *23,* 149-170.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bologna Declaration. *The European Higher Education Area. Joint declaration of the European Ministers of Education.* Bologna Process. Convened in Bologna, 1999, June 19.

Brown, J. D. (2013). Research on computers in language testing: Past, present and future. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary Computer-Assisted Language Learning* (pp. 73-94). London: Continuum.

Carr, N. (2011). *Designing and Analyzing Language Tests: A Hands-on Introduction to Language Testing Theory and Practice.* Oxford: Oxford University Press.

Celce-Murcia, M. (Ed.). *Teaching English as a Second or Foreign Language* (3rd ed.) (pp. 13-28). Boston, Massachusetts: Heinle & Heinle.

Chapelle, C. A. (2001): *Computer Applications in Second Language Acquisition*. New York, Cambridge University Press.

Chapelle, C. & Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.

Council of Europe. (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Díez-Bedmar, M. B. (2011a). Spanish pre-university students' use of English: CEA results from the university entrance examination. *International Journal of English Studies*, *11*(2), 141-158.

Díez-Bedmar, M. B. (2001b). The English Exam in the University Entrance Examination: An Overview of Studies. *Revista Canaria de Estudios Ingleses*, *63*, 101-112.

European Commission. (2012). *First European Survey on Language Competencies: Final Report*. Brussels: European Commission.

Eurostat News Release. (2013). Two-thirds of working age adults in the EU28 in 2011 state they know a foreign language. [Press release]. Retrieved from http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-26092013-AP/EN/3-26092013-AP-EN.PDF

Fernández Álvarez, M. (2007). *Propuesta metodológica para la creación de un nuevo examen de inglés en las pruebas de acceso a la universidad.* Unpublished PhD dissertation. University of Granada, Spain.

Fernández Álvarez, M. & García Laborda, J. (2011). Teachers' interest for a computer EFL University Entrance Examination. *British Journal of Educational Technology, 46*(6), 136-140.

Fernández Álvarez, M., & Sanz Sainz, I. (2005). Breve historia del examen de Selectividad. In H. Herrera Soler, & J. García Laborda (Eds.), *Estudios y criterios para una Selectividad de calidad en el examen de Inglés* (pp. 19-26). Valencia: Editorial Universidad Politécnica de Valencia.

García Laborda, J. (2004): HIEO: Investigación y Desarrollo de una Herramienta Informática de Evaluación Oral multilingüe. *Didáctica (Lengua y Literatura), 16,* 77-88.

García Laborda, J. (2006). ¿Que pueden aportar las nuevas tecnologías al examen de selectividad de inglés?: un análisis de fortalezas y oportunidades. *Revista de Ciencias de la Educación, 206,* 151-166.

García Laborda, J. (2007). On the net: Introducing standardized EFL/ ESL exams. *Computers & Education*, *11*(2), 3-9.

García Laborda, J. (2010). ¿Necesitan las universidades españolas una prueba de acceso informatizada? El caso de la definición del constructo y la previsión del efecto en la enseñanza para idiomas extranjeros". *Revista de Orientación y Psicopedagogía, 21*(1),71-80.

García Laborda, J., Bakieva, M., González Such, J., & Sevilla Pavón, A. (2010) Item transformation for computer assisted language testing: The adaptation of the Spanish University entrance examination. *Procedia - Social and Behavioral Sciences, 2*(2), 3586-3590.

García Laborda, J. & Gimeno Sanz, A. (2008). Adaptación del examen de inglés de las pruebas de acceso a la universidad a un entorno informático: Estudio sobre la tipología de preguntas. *Proceedings of the XXV Congreso Nacional de Lingüística Aplicada*, Murcia, Spain (pp. 723-730). Murcia: Servicio de publicaciones de la Universidad de Murcia.

García Laborda, J. & Fernández Álvarez, M. (2010). Variables sexo, edad y lugar de trabajo en las actitudes de los profesores hacia la interacción oral en L1 y L2 en la clase de inglés de segundo de Bachillerato, *Porta Linguarum 14,* 91-103.

García Laborda, J. & Fernández Álvarez, M. (2012). Actitudes de los profesores de Bachillerato de Alcalá y Navarra ante la preparación y efecto de la PAU. *Revista de Educación, 357,* 29-54.

García Laborda, J., Giménez López, J.L. & Magal Royo, T. (2011). Validating mobile devices in the Spanish University Entrance Exam English paper. *The New Educational Review*, *25*(3), 160.

García Laborda, J., Magal Royo, T. & Enríquez Carrasco, E.V. (2010). Teachers' trialing procedures for Computer Assisted Language Testing Implementation. *Eurasian Journal of Educational Research, 39,* 161-174.

García Laborda, J., & Magal Royo, T. (2007). Diseño y validación de la plataforma PLEVALEX como respuesta a los retos de diseño de exámenes de lenguas para fines específicos. *Ibérica, 14,* 79-98.

García Laborda, J., & Magal Royo, T. (2009) Training senior teachers in compulsory computer based language tests. *Procedia - Social and Behavioral Sciences*, *1*(1), 2009, 141-144.

Godwin-Jones, R. (2011). Mobile apps for language learning. *Language Learning & Technology, 15*(2), 2-11.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177-182.

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning, 25*(3), 257-275.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing, 29*(2), 291-308.

Kukulska-Hulme, A. & L. Shield (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction, *ReCALL*, 20 (3). 271–289.

Lim, G. (2013). Components of an elaborated approach to test validation. *Cambridge English: Research notes*, *51*, 11-14.

Mcnamara, Tim. (2000). *Language Testing*. Oxford: Oxford University Press.

Martín-Monje, E. (2012a): La nueva prueba oral en el examen de inglés de la Prueba de Acceso a la Universidad. Una propuesta metodológica, *Revista de educación*, *357*, 143-161.

Martin-Monje, E. (2012b). The present and future of teaching computer-assisted language: The end of an era? *Revista De Lingüística y Lenguas Aplicadas*, *7*, 203-212.

Martínez Sáez, A., Sevilla Pavón, A., García Laborda, J. & Enríquez Carrasco, E. (2011). Retos y propuestas ante la inminente implantación de las destrezas orales en el examen de lengua extranjera en la futura PAU. *Didáctica de la Lengua y Literatura*, *23*, 321-329.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testi*ng, *13*, 241-256.

Newman, J. M. (Ed.). (1985). *Whole Language Theory in Use*. London: Heinemann.

Norris, S. P., Leighton, J. P., & Phillips, L. M. (2004). What is at stake in knowing the content and capabilities of children's minds?: A case for basing high stakes tests on cognitive models. *Theory and Research in Education*, *2*(3), 283-308.

Oller, J.W. (1979). *Language Tests at School. A Pragmatic Approach*. Harlow (UK): Longman.

Oller, J.W., 1983: *Issues in Language testing Research*, Rowley (Massachusetts): Newbury House.

Richards, J.C. & Rodgers, T.S. (2001). *Approaches and methods in language Teaching*. New York: Cambridge University Press. Sauvignon, Sandra, 2001 [1979]. Communicative language teaching for the twenty-first century” in M. Celce-Murcia (Ed.), *Teaching English as a Second or Foreign Language* (pp. 13-28). Boston, Massachusetts: Heinle & Heinle.

Poehner, M. E. (2007). Beyond the test: L2 Dynamic Assessment and the transcendence of mediated learning. *The Modern Language Journal*, *91*, 323-340.

Stahl, J., Bergstrom, B., & Gershon, R. (2000). CAT administration of language placement examinations. *Journal of Applied Measurement*, *1*(3), 292-302.

Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, *42*(1), 1-40.

Talaván, N. (2010). Claves para comprender la destreza de la comprensión oral en lengua extranjera. *Epos*, *26*, (198-216).

US. Department of Education (2010). *Transforming American Education. Learning Powered by Technology. National Education Technology Plan*. Alexandria (VA): Education Publications Center.

Van Compernolle, R. A. (2011). Responding to questions and L2 learner interactional competence during language proficiency interviews: A microanalytic study with pedagogical implications. In J. K. Hall, J. Hellermann, & S. Pekarek Doehler (Eds.), *L2 competence and development* (pp. 117-144). Bristol: Multilingual Matters.

Vida, J. (2013, June 14). El Govern investigará los errores en los exámenes de selectividad. *La Vanguardia*. Retrieved from http://www.lavanguardia.com/vida/20130614/54375637725/govern-investigara-errores-selectividad.html

Weir, C. (2013). Measured constructs: A history of Cambridge English language examinations 1913-2012. *Cambridge English: Research Notes*, *51*, 2-7.

Weinberg, A. (2001). Comparaison de deux versions d'une test de classement: Version papier-crayon et version informatisee (comparison of two versions of a placement test: Paper-pencil version and computer-based version). *Canadian Modern Language Review*, *57*(4), 607-627.

Xiao-Fan, L. (2000). Computer-assisted English test item analysis. *Computer Assisted Language Learning*, *13*(1), 43-48.

Zahedi, K., & Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): Evidence from Iranian test-takers. *Educational Assessment, Evaluation and Accountability*, *24*(3), 263-277.