Choosing a speaking test in English as a foreign language for the university entrance exam

Marcos PEÑATE CABRERA
Universidad de Las Palmas de Gran Canaria
mpenate@dde.ulpgc.es

Recibido: junio 2013
Aceptado: julio 2014

ABSTRACT
Language assessments are valuable tools to provide information such as whether individual students are ready to move on to another unit of instruction, in this case the university. In the research outlined in this article, our main aim was to develop a suitable speaking exam. We began by studying and designing three different types of tests, combining different formats and resources: one-to-one interview based on a photo, one-to-one interview based on a comic strip and a dialogue in pairs. After piloting two examples for each type of test, we chose the most suitable using the level of difficulty and discrimination power of each item. Then we implemented the three types of test to a statistical significant sample (603 subjects) of the population of students in their final year of secondary education in the Canary Islands. Finally we used the *Testfact 4.0* program to analyze the reliability and level of difficulty of each test, as well as the level of difficulty and discrimination power of each item. This study allowed us to establish the strengths and weaknesses of the different types of tasks used.
Keywords: Language oral test, university entrance exam, reliability, level of difficulty, item discrimination power.

Selección de un test para la evaluación de la expresión oral en lengua extranjera en la prueba de acceso a la universidad

RESUMEN
En la investigación que aquí presentamos, se estableció como objetivo desarrollar y analizar diferentes formatos de evaluación de la expresión oral en la lengua extranjera (inglés) para la prueba de acceso a la universidad. Para dicho fin se estudiaron y diseñaron tres tipos de tests combinando diferentes formatos y recursos: entrevista individual a partir de una foto, entrevista individual a partir de una tira cómica y diálogo en parejas. Tras pilotar dos pruebas para cada tipo de test, se eligió la más adecuada para cada uno de ellos utilizando para tal fin el nivel de dificultad y de discriminación de cada ítem. Las tres pruebas seleccionadas fueron utilizadas para evaluar una muestra estadísticamente representativa (603 sujetos) del conjunto de la población de estudiantes en su último año de bachillerato en Canarias. Con el programa estadístico *Testfact 4.0* se analizó la fiabilidad y nivel de dificultad global de cada prueba, además del nivel de dificultad y de discriminación de cada

ISSN: 1130-0531
http://dx.doi.org/10.5209/rev_DIDA.2014.v26.46849

ítem. Este estudio nos permitió establecer las debilidades y fortalezas de los diferentes tipos de tareas utilizadas.
Palabras clave: Evaluación oral, examen de acceso a la universidad, fiabilidad, nivel de dificultad, nivel de discriminación de los ítems.

Choix d'un test pour l'évaluation de l'expression orale en langue étrangère dans les épreuves d'accès à l'université

RÉSUMÉ
Dans l'étude que nous présentons, l'objectif était de développer et d'analyser les différents formats d'évaluation de l'expression orale en langue étrangère (anglais) pour l'épreuve d'accès à l'université. Pour y arriver, nous avons examiné et élaboré trois types de tests combinant des formats et des supports différents : l'entretien professeur/élève à partir d'une photo, l'entretien à partir d'une bande dessinée, et le dialogue par paires sur consignes spécifiques. Nous avons mené une expérience pilote avec deux modèles d'épreuves pour chaque type de test ; après cette application, nous avons choisi la plus adaptée pour chaque test, en nous appuyant sur le niveau de difficulté et de discrimination de chaque item. Les trois épreuves calibrées ont été employées pour évaluer un échantillon représentatif (603 sujets) de la totalité d'étudiants de la dernière année du baccalauréat aux îles Canaries. Nous avons vérifié avec le logiciel *Testfact 4.0* la fiabilité et le degré de difficulté global de chaque épreuve et, en outre, le niveau de difficulté et de discrimination de chaque item. L'analyse nous a permis d'établir les faiblesses et les points forts des différents types de tâches présentées.
Mots-clés: Évaluation de l'oral, examen d'accès à l'université, fiabilité, degré de difficulté, niveau de discrimination des items.

INTRODUCTION

The university entrance exam in Spain includes, as one of the assessed subjects, English as a foreign language. This specific test has been focused on the formal aspects of the language and, especially, on the reading and writing skills. That is, the oral skills have been completely pushed into the background. However, the Spanish educational authorities have decided on a change of direction in the assessment of the foreign language, incorporating the listening and speaking skills, as was described in this journal by Martínez et al. (2011). Nevertheless, this decision is not free of challenges and considerations that will imply an enormous effort to carry it out.

In his introductory article to the monograph on this subject in the Revista de Educación (nº 357), García Laborda (2012, 18) reviews the tasks that need to be carried out, among which we emphasize the following: the research needs to be updated, the prospective implications in the classroom and in students themselves need to be analyzed and the test construct and the issue of delivery need to be addressed.

To assess the listening skill, classrooms with adequate acoustic means will be needed so that students may have the right conditions to carry out the tasks required in the listening activities or incorporate its assessment as part of the speaking exam (Wood and Bobb, 2012, 108)

Incorporating oral tests implies that test developers will need to address questions about costs and availability of resources for collecting this information. To reduce costs and the amount of time needed to assess large numbers of students, some researchers are designing tools to carry out a computer-based oral assessment (García Laborda, 2004 and 2006; Magal-Royo and Giménez López, 2012; Martín-Monje, 2012) and even mobile-based testing (García Laborda, Magal Royo, Litzler and Giménez López, 2014). But besides considerations related to the exam organization (practicality), it has to be established what type of test is the most suitable to assess the speaking ability of the students (validity), specifying at the same time the level of reliability of the chosen test.

In the Canary Islands, where oral expression has been assessed at each one of the external assessments conducted by the Instituto Canario de Evaluación (Peñate and Bazo, 2007), we have attempted to throw some light on the concepts of validity and reliability. It is for this reason that Wood and Bobb (*Op.cit.*), the university coordinators, have been working on what to include, how the exam could be run and how to assess the students' performance. To further develop the previous work, what we put forward here is a statistical study on three different kinds of test: one-to-one interview based on a photo, one-to-one interview based on a comic strip and a dialogue in pairs.

1. STATE OF THE ART

In this section we are going to revise some articles dealing with the following two topics: paired testing versus the one-to-one format and types of tasks traditionally used in oral assessment.

1.1. Paired testing versus the one-to-one format

Research on the one-to-one format (Ross and Berwick, 1992; Young and Milanovic, 1992) has shown that the interaction in this type of speaking test was asymmetrical in terms of contingence and goal-orientation, but above all because of the unequal status of the two participants: interviewer and test taker. Hughes (1989, 104) insists on this important feature and points out at least one potentially serious drawback of the traditional one-to-one interview format: the power

relationship which exists between tester and candidate. So it seems that the fixed role relationship in a one-to-one test makes it difficult for the candidate to escape this asymmetry. More recent studies dealing with the one-to-one model have investigated the perceptions of the verbal, paralinguistic and nonverbal discourse behaviors of the examiner. So in the ethnographic research carried out by Plough and Bogart (2008), test takers found these behaviors to be meaningful in terms of their comfort level during the test.

And this way of thinking was already taking shape in the 90s as can be stated in the articles of Foot (1999) and Saville and Hargreaves (1999). But this new approach was again the source of differing points of view and so we can see, for example, that Saville and Hargreaves (*Op.cit.*) emphasized the advantages of the paired format, especially the relaxed atmosphere, the variety of language produced and the washback effect in the classroom. Whereas Foot (*Op.cit.*, 36) though admits that this new procedure has a number of advantages, in particular that of helping candidates to feel relaxed, he regrets that although the University of Cambridge Local Examinations Syndicate has ten years' experience of examining candidates in pairs at one level, neither their original research findings, nor the results of any subsequent monitoring of the changes have been published. This criticism concerning the lack of research evidence was replicated two years later in the Research Notes of the above mentioned university by means of an article of Lynda Taylor (2001). In this article, she outlines two internal studies set out to compare the paired and one-to-one speaking test formats. The results obtained in the first quantitative study suggested that the paired format generated a much richer and more varied sample of spoken language than in the one-to-one format. And the second, more qualitative study, used an observation checklist to analyse the distribution of certain speech functions, identifying a total of thirty communicative language functions in the paired format. Also in the year 2001 we can find another article that is a direct reaction to the problems pointed out by Foot (*Op.cit.*). Együd and Glover (2001, 72) argue in favor of the paired format providing qualitative date of the paired oral testing carried out in a secondary school context. By giving examples of the students' opinions about this type of test, they prove that pairings can be just as successful (linguistically and emotionally) with school students. From the same year, there is an article by Swain (2001, 295) where she considers the implications of the paired format from the testing point of view and indicates that serious thought needs to be given to the most adequate and fair means of scoring the resulting linguistic activity.

In the following years there has been a considerable amount of research, mainly case studies that have pinpointed specific features related to the paired format. So according to the data presented in different studies, pairing potentially affects linguistic performance if one candidate has higher linguistic ability than the other, and also seems to affect the amount of talk produced (Norton, 2005). More recently, Davis (2009) has examined the influence of interlocutor proficiency and

was able to determine that it had no observable effect on ability measures, but lower-level examinees produced more language (words) when working with a higher-level partner. In other cases the main research aim is not the amount but the quality of talk, that is, the nature of talk elicited by paired format. Galaczi's analysis (2008) highlights global patterns of interaction termed Collaborative, Parallel and Asymmetric and salient features of interaction characteristic to each pattern. O'Sullivan (2002) explored the effect on pair-task performance of test takers' familiarity with their partner and found that subjects achieved higher scores when working with a friend. In another study carried out ten years later (Chambers, Galaczi and Gilbert, 2012), the comparative analysis of the scores awarded to the pairs been tested with a friend and to the pairs of unknown people indicated small but not meaningful differences in overall speaking test performance and performance by assessment criteria. From a different perspective, Ockey (2009) has investigated the effects of group members' personalities and the findings of the study suggest that when the group oral is used, raters should be careful specially in assigning scores that are not based on a comparison of proficiencies of group members.

But there have also been studies that have adopted an experimental approach and have compared the two test formats. Such is the case of the study of Brooks (2009) where she examines the interaction of adult ESL test-takers in two tests of oral proficiency: one in which they interacted with an examiner (the individual format) and one in which they interacted with another student (the paired format). The findings from the quantitative analyses carried out by Brooks show that the test-takers performed better in the paired format being their scores on average higher than when they interacted with an examiner. Similar positive results were found by means of the qualitative analysis of the test-takers' interaction with other students in the paired test. This interaction was much more complex as there was a bigger variety of interactive features in their dialogue and a more balanced distribution of those features between the test-takers (Brooks, *Op.cit.*, 360-1).

1.2. Types of tasks

McNamara (1996) sees test performance as being affected by a number of factors related to the test-taker and the interlocutor but also to the task. As interest in the use of tasks has grown, one of the research fields has focused on the types of tasks, processes and abilities to be assessed (Stoynoff, 2012, 524). The popularity of performance testing resulted in a growing interest in tasks as a vehicle for assessing learner ability. Task-based assessment requires the test taker (Elder, Iwashita and McNamara, 2002, 347) to engage in the performance of tasks which simulate the language demands of the real world situation with the aim of eliciting an authentic sample of language from the candidate. The properties of such tasks and the influence of these properties on learner performance are now being widely researched, with some scholars focusing on strengthening the links between test

tasks and their real world counterparts (Bachman and Palmer, 1996; Douglas, 2000) and others on the effect on candidate production of manipulating different task characteristics in the test situation (Elder et al., *Op.cit.*; Norris, Brown, Hudson and Bonk, 2002). There have also been a lot of theoretical studies trying to establish the factors that contribute to task complexity. One example of these frameworks is the one proposed by Robinson (2001). The author distinguishes between resource-directing factors (number of task elements, reasoning demands, immediacy of information provided) and resource-depleting factors (planning time, number of tasks, prior knowledge). And this theoretical approach has become a sort of springboard in the designing of further research dealing with topics like the difficulty of tasks in their relation with a specific first language cultural background (Fulcher and Márquez Reiter, 2003) or the amount of planning time available (Wigglesworth and Elder, 2010).

Thus, despite the fact that language testers have been researching the relationships among task characteristics and task difficulty for over a decade, the results of this research seem to have brought us no closer to an understanding of this relationship (Bachman, 2002, 463)

To find a solution to this problem, we decided to start studying the different types of tasks used in some of the most popular speaking exams and trying to evaluate their validity. Here we are referring to their face validity. That is, activities that can be considered useful and not merely oral tasks without any relationship with real language exchange. Moreover, if we choose unsuitable activities we would be sending the wrong message to teachers, as they will adapt their teaching methodology and lesson content to reflect the test's demands (Taylor, 2005, 154). So validity is expected to have a double effect: on students' motivation and the right washback effect on teachers (Amengual-Pizarro, 2009, 593).

Our first decision was to model our test on the high-stakes EFL tests that are widely used. For doing that we have always kept in mind at least two questions (Bachman and Palmer, 2010, 5): Is our situation similar enough to the ones for which these high-stakes tests were developed to make them appropriate? Are the abilities tested in those tests the ones we need to test?

The Test of English as a Foreign Language (TOEFL) measures the speaking skill in about 20 minutes by means of two questions to express an opinion on a familiar topic, based on the students' personal knowledge and experience, and four integrated speaking tasks based on what is read or listened to. For the first two questions, students have a short amount of time after they read each question to prepare their response. For tasks 1 and 2, test takers read a short passage and then they hear a short talk on the same subject. Then they will answer a question that relates to both of them. In tasks 3 and 4 students hear a conversation and a lecture and then they are asked a question about each. Cumming, Grant, Mulcahy-Ernt and Powers (2004) designed a study to explore how some ESL teachers evaluated some prototype tasks for the TOEFL, judged the authenticity of their students'

performance using the given tasks and, finally evaluated whether they fulfilled the purposes for which they had been designed. Of the seven tasks used, three were for the speaking skill (in response to a lecture, to a conversation and to a reading passage). The teachers rated the prototype speaking tasks positively overall, though they also indicated that students with lower proficiency in English were hampered in their speaking performance if they did not comprehend the ideas, vocabulary, or background context of the reading or listening stimulus texts (Cumming et al., *Op.cit.*, 134-5).

Of the University of Cambridge ESOL examinations we are going to focus on the Preliminary English Test (PET) as this exam is related to level B1 of the Common European Framework of Reference for the Languages, a level that is expected to be developed, though not reached, in the final years of Spanish secondary education. The PET assesses the speaking skill in about 10-12 minutes. The assessment is conducted face-to-face with another candidate, as it intends to make it more realistic and reliable. There are two examiners. One of the examiners is the interviewer and the other one the rater. The test has four parts. In part 1, the examiner asks personal related questions, whereas in part 2 the examiner gives the candidate some pictures and describes a situation and afterwards test takers are expected to talk to the other candidate and decide what would be best in the given situation. Then the examiner hands out a colour photograph and the student has to talk about it. Finally in task 4, both test takers are expected to discuss about the topic presented by means of the picture presented previously.

Trinity Graded Examinations in Spoken English (GESE) has 12 graded tests in spoken English that are organized in four development stages (initial, elementary, intermediate and advanced). As grades 5 and 6 of the elementary stage are associated to the B1 level, we will concentrate on the way they are evaluated. The examination procedure at this level has two main parts: discussion of a prepared topic by the student and a conversation with the interviewer on two randomly selected subject areas of those listed for the grade. Besides giving information, making statement and responding, test takers are expected to ask at least one question to the examiner.

To conclude this section on the types of tasks, we must mention that the two types of tasks which seem to be most widely agreed on at present, are those supported by Amengual-Pizarro and Méndez (2012, 117): the performance of a sustained monologue and the development of a conversation between two or three candidates. The agreement on these two types of tasks springs from the need to assess the speaking skill (spoken production) and that of spoken interaction proposed by the Common European Framework of Reference for Languages (CEFR) of the Council of Europe (2001). Although there are researchers who suggest, and we believe rightly so, the need to study what type of tasks are carried out in the classroom and to try to include them in the tests we carry out ourselves (García Laborda and Fernández Álvarez, 2012, 33).

2. METHODOLOGY

The concept of reliability has been one of the main factors in modern language testing and it has obviously been used in studies dealing with the University Entrance Examination. This is the case of Herrera Soler (2001) and Amengual-Pizarro (2006) who studies the effect of raters' gender and working place on reliability. However, it has been dealt with from different perspectives. So at the beginning of this century, the classical norm-referenced reliability coefficients was pushed into the background as this type of test compares a student's test performance with that of a sample of similar students. Instead, researchers have opted for newer and more powerful quantitative methodologies (Bachman, 2000, 4). Among these quantitative methodologies, criterion-referenced measurement, generalizability theory, structural equation modeling and item response theory have been the most used in this research field. Criterion-referenced measurement (CRM) focuses on individual, differentiated assessment which describes the type of behavior expected of a person with a given score; Generalizability theory (G-theory) is based on the simultaneous analysis of multiple sources of measurement error; Structural equation modeling (SEM) is a statistical technique for studying causal relations among variables using a combination of statistical data and qualitative causal assumptions; and Item response theory (IRT) is a measurement model that enables us to estimate the statistical properties of items and the abilities of test takers.

Since the main purpose of our study is to estimate the statistical properties of items and the abilities of test takers, Item Response Theory (IRT) will be the methodology that we will adopt. IRT is based on the fundamental theorem that an individual's expected performance on a particular item is the result of two complementary features: the level of difficulty of the item and the individual's level of ability (Bachman, 1990, 203).

2.1. Types of tests designed

Once we finished studying all the possibilities outlined by the main speaking tests used in our context, we began designing the types of tests we wanted to implement, and in this decision-making process we were specifically interested in the relevance and representativeness of test content, the appropriateness of task design and rating scales.

For our study, we chose three different types of tests that were piloted in a previous study (Wood, Peñate and Bazo, 2007). Tests 1 and 2 (see Appendix) are implemented by means of an interview conducted by the rater and students are evaluated individually. In the first type, pupils are presented with a picture and they have to answer five questions directly related to it, then five more questions only partially related to the picture, and finally five personal related questions, although based on the topic depicted in the picture. Test 2 is similar to the previous one, with the only exception that we used a comic strip instead of a picture. The main

difference between using a picture and a comic is that, by choosing the latter we are dealing with the two skills of the CEFR mentioned above: speaking versus spoken interaction. With the comic we are assessing students' ability to describe events specifically by means of items 2 to 5 (What happens in strip one / two / three / four?) This ability is labeled as "Sustained monologue" and is considered to be part of the speaking skill. Whereas the ability to maintain a conversation is part of the spoken interaction skill and is assessed in the remaining items.

The third test (see also Appendix) is completely different as now the evaluation is carried out with two students simultaneously. Before the test each student is given a card with five prompts indicating the information they have to obtain from their partners. After a few minutes, each student has to ask five questions (items 1 to 5), and answer the questions of his or her partner (items 6 to 10). Once the students have finished asking and answering questions, the rater will ask five questions to be answered by both students (items 11 to 15). In the three tests, there are five final items by means of which the rater gives an overall assessment.

As for the rating scales, we used the following one for the first 15 items: 0 failure to reply; 1 incorrect answer; 2 partially correct answer; and 3 correct answer. For the overall assessment items we designed five rubrics for use of lexis, speech, grammar, interaction and pronunciation. Each rubric had four detailed descriptors that were labeled in the same way as the previous one: from 0 to 3.

2.2. Subjects

As the main purpose of this research was to study the effects of different types of tests for the oral assessment at the university entrance exam, we conducted this study with the students in their final year of secondary education in the Canary Islands. Taking into account the population (around 2600 students), it was established that our statistical significant sample should be of 600 subjects as we were selecting groups of students and not individual students randomly selected from the whole population. To reach this number, 24 secondary schools were randomly chosen and in each school one group was again randomly selected. Each type of test was implemented in eight secondary schools as can be seen in the following Table. The 24 chosen schools were on the two main islands: Gran Canaria and Tenerife (12 on each). The selection included state and public schools, as well as urban and rural ones.

	Schools	Raters	Students	
			Expected	Assessed
Test 1	8	4	204	195
Test 2	8	4	204	192
Test 3	8	4	231	216

Table 1: Sample population

As can be seen in the above Table, the total number of students assessed was of 603, so we reached the number of students needed to have the significant sample of population mentioned previously (600).

3. RESULTS

For the statistical analysis we used Testfact 4.0, one of the software programs published by Scientific Software International, as it performs classical test scoring, item analyses, and item factor analysis. This program was designed originally for a national testing service and has all of the features needed for processing data from binary scored tests, subtests, or scales. Since our rating scales were divided into four categories, we had to dichotomize those scales into 0 and 1 (error and correct answer) and by doing so we were able to use the above mentioned program to carry out the analysis.

The results of the analysis will be presented in two sections. First we want to study the reliability and the level of difficulty of each test. The second section will be devoted to the item analysis, paying special attention to the level of difficulty and discrimination power of each one of them.

3.1. Reliability and level of difficulty of each test

3.1.1. Reliability of each test

Reliability is used to describe the overall consistency of a test. A test is said to have a high reliability if it produces similar results under consistent conditions. To carry out this study we used the Cronbach's alpha coefficient, since it is commonly used as a measure of the internal consistency or reliability of a psychometric test score for a sample of examinees. Besides, it has the great advantage of being considered as an estimate of the lower limit of the reliability coefficient level. This is also a popular method as it does not need a test-retest to check stability, since only a single test is needed for estimating internal consistency.

We should also bear in mind a criterion that will allow us, once we have estimated the reliability coefficients, evaluate if they are adequate to the aim of this study. This coefficient takes values between 0 and 1 such that the larger this quantity, the better the reliability. Generally speaking, the higher the alpha is, the more reliable the test is. There is not a commonly agreed cut-off, though

coefficients at or above 0,800 are often considered sufficiently reliable to make decisions about individuals based on their observed scores, although a higher value is preferred if the decisions have significant consequences (Webb, Shavelson and Haertel, 2007, 81). Consequently, as the main aim of these tests is to assess the linguistic competence level of individual students, we will take the reliability coefficient level of 0,940 for the whole set (items 1-20) and 0,800 for the different sets or groups of items (items: 1-5, 6-10, 11-15 and 16-20).

Once the relevant calculations have been done, we obtained the following alpha values for each test and subtest, which gives us the reliability levels that can be seen in the following Graph and Table:

Graph 1: Test reliability levels

	Items 1-5	**Items 6-10**	**Items 11-15**	**Items 16-20**	**Items 1-20**
Test 1	**0,727**	**0,790**	**0,798**	0,876	0,932
Test 2	0,842	0,822	0,814	0,897	**0,948**
Test 3	0,804	0,799	0,835	0,948	**0,945**

Table 2: Test reliability levels

As can be seen in the previous Graph and Table, tests 2 and 3 reach alpha values above the established lower limit (0,940) to provide relevant information to

evaluate individual scores, though we must also point out that test 1 is quite close to this lower limit.

The reliability level of the three tests for items 16-20 shows that the overall rating scales, shared by the three tests, seem to be quite relevant and accurate. However, it must be pointed out that these reliability levels are higher in test 3, probably due to the type of assessment implemented.

Items 1 to 5 of test 1 have got the lowest reliability levels of all the obtained results. These are the items directly related to a given picture, where students were expected to describe the situation depicted, infer its meaning and use their imagination. As it happens, this is the only set of items of the three tests where it was required to infer the meaning and to give free rein to one's imagination. In the same way, sets of items 6-10 and 10-15 of test 1 also have reliability levels slightly below the lower limit for a subtest (0,800).

Then we implemented the Spearman–Brown prediction formula, also known as the Spearman–Brown prophecy formula, which is a formula relating psychometric reliability to test length and used by psychometricians to predict the reliability of a test after changing the test length. In our study we obtained the following lengths for our tests.

	N
Test 1	23
Test 2	18
Test 3	19

Table 3: Estimated test lengths

As the number of total items of a tests is one of the factors that have an influence on the reliability, test 1 could be improved by increasing the total number up to 23, whereas test 2 and 3 could reduce the total number of items by two and one item respectively without compromising the reliability levels. So the total number of items planned when we started drafting the tests has turned out to be acceptable, bearing in mind the small modifications outlined above.

So far we have verified that test 1 seems to be the weakest. On this matter, it should be noted that this type of test is probably the most frequently used to assess the speaking skill in this type of assessment and, precisely in this test 1, the set of items with the lowest reliability level (items 1-5) is its main distinctive feature.

We can sum up this section devoted to the reliability of each test, stating that the three tests have the required coefficient levels and only test 1 require some small improvements, particularly the inclusion of three extra items. And finally we have to point out that test 2 have proved to be the most reliable.

3.1.2. Level of difficulty of each test

The next step of our analysis was to study the distribution range of the scores obtained in each test with the final objective of establishing the level of difficulty of each one.

Since each test has 20 items assessed by a rating scale that goes from 0 to 3, the total score of each test will be in a scale from 0 to 60 points (20 items * 3 = 60).

	Rank			**Normal distribution**		
	Minimum	**Maximum**	**Average**	**Lower limit**	**Upper limit**	**% with scores > = 30**
Test 1	2	60	39,8	26,0	53,5	75,4
Test 2	1	60	34,1	19,1	49,2	61,5
Test 3	0	60	38,2	25,0	51,4	72,0

Table 4: Level of difficulty of each test

The average score of each test goes from 34,1 points of test 2 to 39,8 points of test 1. Therefore, test 1 turned out to be the easiest one, being the most difficult one test 2, though the differences are not very big. The percentage of test-takers with scores of or higher than 30 (half of the maximum possible score) ranges from 61,5% in test 2 to 75,4% in test 1.

Here we should not forget that, since the main purpose of these tests is to assess the linguistic oral competence of students individually and not to select candidates, it would not be advisable to have too low scoring rates.

Finally, we can see that the changes implemented after piloting these tests have allowed us to have acceptable difficulty levels. Test 1 and 3 have similar averages, 39,8 and 38,2 respectively, whereas test 2 has 34,1, which means an average slower in at least four points. This average difference, would be on a scale 0-10 of 0,68 and 0,95. In the same way, when comparing the number of test-takers that obtain at least half of the assigned score (60), a difference of at least 10,5% can be seen.

3.2. Item analysis: level of difficulty and discrimination power

This section is devoted to the analysis of each item from a double perspective: its level of difficulty and its discrimination power. By doing it, we will be able to know the capacity of each item to distinguish between high and low level students in each of the tests.

3.2.1. Level of difficulty per item

The difficulty of an item is understood as the proportion of the persons who answer a test item correctly. To calculate the level of difficulty of the items we have divided the mean score of each one of them by three, obtaining so the

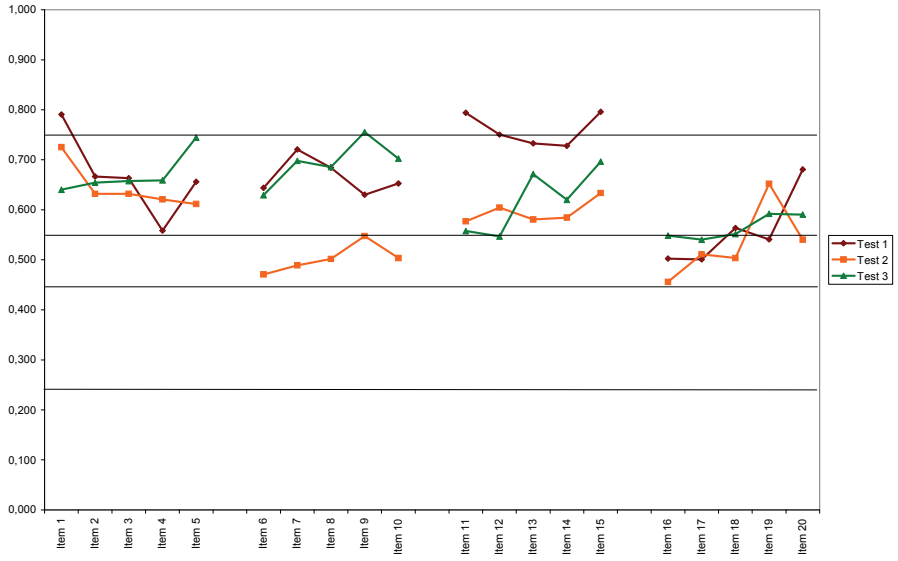
difficulty level (DL) measured in a scale from 0 (Maximum difficulty) to 1 (Minimum difficulty). What this means is that it has to do with an inverse relationship: the greater the difficulty of an item, the lower its index.

		% distribution of items per difficulty level		
Category	**Criterion**	**Test 1**	**Test 2**	**Test 3**
Very difficult	DL below 0,25			
Difficult	DL between 0,25 and 0,44			
Normal	DL between 0,45 and 0,54	15,0%	45,0%	15,0%
Easy	DL between 0,55 and 0,74	65,0%	55,0%	80,0%
Very easy	DL above 0,74	20,0%		5,0%

Table 5: Distribution of items per difficulty level for each test

As can be seen on the above Table, the percentage of easy and very easy items is of 85% in tests 1 and 3, whereas in test 2 that percentage is reduced to only 55%. In the same way, it should be noted that in the 85% of test 1, 20% of those items are considered to be very easy, however this percentage is only 5% in test 3, which seems a more reasonable distribution. Test 2 does not have a balanced distribution as the percentage of very easy items is 0% and also it does not seem reasonable to have a similar distribution of normal and easy items.

Taking as a difficulty criterion the categories proposed by Pérez Juste and García Ramos (1989, 306), it can be established that none of the items in the tree tests turned out to be difficult or very difficult for the test takers, between 15% and 45% of those items depending on the test had a medium level of difficulty, between 55% and 80% of items can be considered as easy, and a 20% of test 1 and a 5% of test 3 proved to be very easy.



Graph 2: Difficulty level per item

Test 1	**0,791**	**0,667**	**0,663**	**0,558**	**0,656**	**0,644**	**0,721**	**0,684**	**0,630**	**0,653**	**0,794**	**0,750**	**0,733**	**0,728**	**0,796**	0,503	0,501	**0,564**	0,541	**0,681**
Test 2	**0,725**	**0,632**	**0,632**	**0,621**	**0,612**	0,471	0,489	0,502	0,548	0,504	**0,577**	**0,604**	**0,581**	**0,584**	**0,634**	0,456	0,511	0,504	**0,652**	0,540
Test 3	**0,640**	**0,654**	**0,657**	**0,659**	**0,745**	**0,629**	**0,698**	**0,685**	**0,755**	**0,702**	**0,558**	0,547	**0,671**	**0,620**	**0,696**	0,548	0,540	**0,551**	**0,592**	**0,590**

Table 6: Difficulty level per item

Of the items that turned out to be very easy (DL >= 0,750), four belong to test 1 and one to test 3. On the other hand, none of the items have a difficult level lower than 0,450.

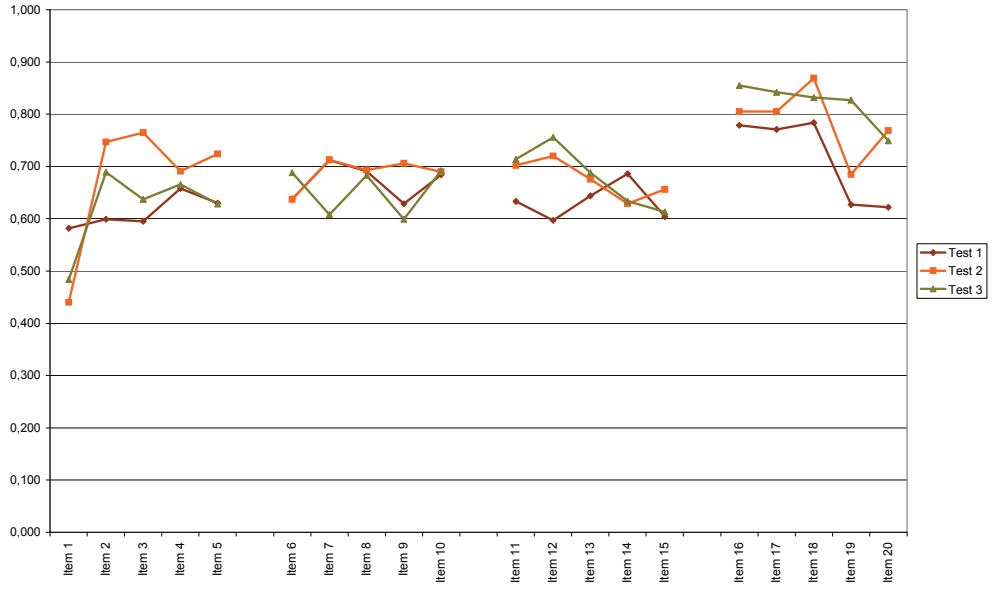
Looking carefully at the above Graph depicting the difficulty level per item, we can see that the lower level of difficulty of test 1 is due above all to the group of items from 11 to 15, whereas the higher level of difficulty in test 2 is due especially to the group of items from 6 to 10.

Items 11 to 15 in test 1 are related to questions about students' personal experiences, as they ask test takers to talk about their free time activities and besides the interviewer's questions could be answered with short direct sentences. The same applies to item 9 of test 3 where students were asked about their summer sport and where they play it.

In items 6 to 10 in test 2, students are required to put themselves in the shoes of people who have nothing to do with them (people at a court) which requires a higher level of abstraction. In the same way, item 12 of test 3 (*What do your parents or friends tell you about your style?*) asks the test taker to express the opinion of others about a very specific topic. In the rest of the items, the three tests behave in a similar way.

3.2.2. Item discrimination power

As for the discrimination power level of each item, most of them have obtained high values (of 0,600 or higher) as can be seen in the Graph below. Item 1 in the three tests gets the lowest level, being the only one below 0,600. In the case of test 1, item 1 is one of the four easiest items in the whole oral exam, and the same applies to test 2, as item 1 is the easiest one of the whole test as we saw in the previous section devoted to the level of difficulty of each item. Here we must bear in mind that the discrimination power of an item is reduced if the level of difficulty is too slow or too high: that is, if most of the test takers get the item right or wrong, its discrimination power is reduced.



Graph 3: Item discrimination power

None of the items have obtained a discrimination level lower than 0,400. So, according to the criterion stated by Suárez Falcón (2006, 403) it can be asserted that all items have an acceptable discrimination power. As was pointed out above, most items have got discrimination levels ranging between 0,600 and 0,800.

On the whole, we can see that the set of items with the highest discrimination levels in the three tests is the last one, that is, the set for items 16-20. We must also remember that this set of items reached also the highest levels of reliability and that they assess the vocabulary, speech, grammar, interaction and pronunciation.

4. DISCUSSION AND CONCLUSIONS

To avoid falling into the temptation to choose the best test of the three types we have tested, we have tried not to forget the following quotation from Bachman and Palmer (2010, 6): "… there is no such thing as the one "best" test, even for a specific situation, and that the terms "good" and "bad" are not very useful for describing a language test. In any situation, there will be a number of alternatives, each with advantages and disadvantages." So let us consider the advantages and disadvantages of the tests according to the results obtained.

We must begin by explaining why we have not implemented the tasks used by TOEFL where test takers are expected to answer some questions after reading a short passage and listening to a lecture. Since we are trying to evaluate the effectiveness of different tasks and contexts for evaluating the oral production of our students, we decided not to use tasks based on listening and reading documents as these could hamper their speaking performance. That is, we wanted to be certain that in case of a poor performance, this was due to the speaking ability of the test taker and not to his or her inability to understand the oral or written text.

Test 1 turned out to be the only one with reliable levels below the established limits. In particular the five questions directly related to the picture obtained the lowest level of the whole study (0,727). According to the Spearman-Brown prediction formula one way of solving this problem is by adding three more items and having a total of 23. When analyzing the level of difficulty of each test, we found that precisely test 1 was the easiest one with an average score of 39,8 of a maximum of 60. The same characteristic comes across when studying the level of difficulty of each item: 65% are easy and 20% very easy.

Test 2 was the most reliable (0,948) but it was also the most difficult one with an average score of only 34,1 out of 60. Considering that the main purpose of the exam is assessing the oral competence of students and not selecting candidates, the level of difficulty is something to be considered. When studying the level of difficulty per item, the previously general feature was confirmed: only 55% of the items are easy and there are not any very easy ones. Precisely 5 out of the 6 most difficult items belong to this test. They are items 6 to 10 where students are asked to express the opinion and feeling of the characters presented in the comic.

Test 3 was quite reliable (0,945) and its level of difficulty is of 38,2, which means that it is easy but it is not the easiest one. The level of difficulty of test 3 can be seen better when looking at the levels of difficulty of the items, 80% are easy and only 5% very easy which seems a more reasonable distribution than the one we saw in test 1. From the perspective of test validity and authenticity, the pairing of

candidates provides a more varied sample of interaction than an individual interview. And as we expect our students to speak to each other in English in realistic situations, so examinations should reflect this.

The rubrics used in the three tests for the overall assessment of the lexis, speech, grammar, interaction and pronunciation (items 16-20) reached the highest reliability levels specially when used for test 3 (0,948) considering that the lower acceptable level for a set of items is of 0,800. Similar results are obtained when studying the discrimination power of each item. Items 16 to 20 of the three tests obtained the highest discrimination levels.

One of the key aims of this project was to improve the impact of the oral exam on teaching. We hope that this test will have the desired effect on the content of secondary school classes, in that much more emphasis will be placed on the teaching of speaking abilities.

Besides, we are convinced that oral assessment will continue to be studied from several different perspectives, which is why, for example, it seems necessary to draw up relation tables (Peñate and Bazo, 2014) that will allow us to develop the needed evaluation standards, but from the perspective of English as an international language (Amengual-Pizarro and Méndez, 2012, 122). Regarding this particular study, the next step will be to do an analysis from a linguistic point of view of the oral production arising from the three tests, using the tests of homogeneity and specificity.

5. BIBLIOGRAPHY

AMENGUAL-PIZARRO, Marian (2006): "Análisis de la prueba de inglés de selectividad de la Universitat de les Illes Balears", in *Ibérica*, 11, 29-59, http://www.aelfe.org/documents/Ib11-03%20Marian%20Amengual%20Pizarro.pdf

AMENGUAL-PIZARRO, Marian (2009): "Does the English Test in the Spanish University Entrance Examination influence the teaching of English?", in *English Studies*, 90, 528-598.

AMENGUAL-PIZARRO, Marian; MÉNDEZ GARCÍA, Mª del Carmen (2012): "Implementing the oral English task in the Spanish University Admission Examination: an international perspective of the language", in *Revista de Educación*, 357, 105-127, http://www.mecd.gob.es/dctm/revista-de-educacion/articulosre357 /re35705.pdf

BACHMAN, Lyle (1990): *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press.

BACHMAN, Lyle (2000): "Modern language testing at the turn of the century: assuring that what we count counts", in *Language Testing,* 17, 1-42.

BACHMAN, Lyle (2002): "Some reflections on task-based language performance assessment", in *Language Testing*, 19, 453-476.

BACHMAN, Lyle; PALMER, Adrian (1996): *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford, Oxford University Press.

BACHMAN, Lyle; PALMER, Adrian (2010): *Language Assessment in Practice*, Oxford, Oxford University Press.

BROOKS, Lindsay (2009): "Interacting in pairs in a test of oral proficiency: Co-constructing a better performance", in *Language Testing*, 26, 341-366.

CHAMBERS, Lucy; GALACZI, Evelina; GILBERT, Sue (2012): "Test taker familiarity and speaking test performance: Does it make a difference?", in *University of Cambridge ESOL Examinations Research Notes*, 49, 33-40.

COUNCIL OF EUROPE (2001): *Common European Framework of Reference for Languages: learning, teaching, assessment*, Cambridge, Cambridge University Press.

CUMMING, Alistair; GRANT, Leslie; MULCAHY-ERNT, Patricia; POWERS, Donald. (2004): "A teacher-verification study of speaking and writing prototype tasks for a new TOEFL", in *Language Testing,* 21, 107-145.

DAVIS, Larry (2009): "The influence of interlocutor proficiency in a paired oral assessment", in *Language Testing*, 26, 367-396.

DOUGLAS, Dan (2000): *Language Testing for Specific Purposes*, Cambridge, Cambridge University Press.

EGYÜD, Györgyi; GLOVER, Philip (2001): "Oral testing in pairs – a secondary school perspective" in *ELT Journal,* 55, 70-76.

ELDER, Catherine; IWASHITA, Noriko; MCNAMARA, Tim (2002): "Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?", in *Language Testing*, 19, 347-368.

FOOT, Michael (1999): "Relaxing in pairs", in *ELT Journal*, 53, 36-41.

FULCHER, Glenn; MÁRQUEZ REITER, Rosina (2003): "Task difficulty in speaking tests", in *Language Testing*, 20, 321-344.

GALACZI, Evelina (2008): "Peer-peer interaction in a speaking test: the case of the First Certificate in English examination", in *Language Assessment Quarterly,* 5, 89-119.

GARCÍA LABORDA, Jesús (2004): "HIEO: Investigación y desarrollo de una herramienta informática de evaluación oral multilingüe", in *Didáctica (Lengua y Literatura),* Publicaciones Universidad Complutense http:///revistas.ucm.es/index.php/DIDA/article/view/DIDA0404110077A/19337, 16, 77-88.

GARCÍA LABORDA, Jesús (2006): "La Plataforma de exámenes multiligüe PLEVALEX: Resultados del diseño y perspectiva de investigación futura de la Plataforma de Exámenes Valenciana de Lenguas Extranjeras", in *Didáctica (Lengua y Literatura),* Publicaciones Universidad Complutense, 18, 135-145, http://revistas.ucm.es/index.php/DIDA/article/view/DIDA0606110135A

GARCÍA LABORDA, Jesús (2012): "Presentación. De la Selectividad a la Prueba de Acceso a la Universidad: pasado, presente y un futuro no muy lejano", in *Revista de Educación*, 357, 17-27, http:// www.mecd.gob.es/dctm/revista-de-educacion/articulosre357/re35701.pdf

GARCÍA LABORDA, Jesús; FERNÁNDEZ ÁLVAREZ, Miguel (2012): "Actitudes de los profesores de Bachillerato adscritos a la Universidad de Alcalá y a la Universidad Pública de Navarra ante la preparación y efecto de la Prueba de Acceso a la Universidad", in *Revista de Educación*, 357, 29-54, http:// www.mecd.gob.es/dctm/revista-de-educacion/articulosre357/re35702.pdf

GARCÍA LABORDA, Jesús; MAGAL ROYO, Teresa; LITZLER, Mary Frances; GIMÉNEZ LÓPEZ, José Luis (2014): "Mobile Phones for Spain's University Entrance Examination Language Test", in *Educational Technology & Society*, 17, 2, 17-30

HERRERA SOLER, Honesto (2001): "The effect of gender and working place of raters on university entrance examination scores", in *Revista Española de Lingüística Aplicada (RESLA)*, 14, 161-179.

HUGHES, Arthur (1989): *Testing for Language Teachers*, Cambridge, Cambridge University Press.

MAGAL-ROYO, Teresa; GIMÉNEZ LÓPEZ, José Luis (2012): "La interactividad multimodal en la sección de lengua extranjera de la Prueba de Acceso a la Universidad en España", in *Revista de Educación*, 357, 163-176, http:// www.mecd.gob.es/dctm/revista-de-educacion/articulosre357/re35708.pdf

MARTIN-MONJE, Elena (2012): "La nueva prueba oral en el examen de Inglés de la Prueba de Acceso a la Universidad: una propuesta metodológica", in *Revista de Educación*, 357, 143-161, http:// www.mecd.gob.es/dctm/revista-de-educacion/articulosre357/re35707.pdf

MARTÍNEZ SÁEZ, Antonio; SEVILLA PAVÓN, Ana; GARCÍA LABORDA, Jesús; ENRÍQUEZ CARRASCO, Emilia (2011): "Retos y propuestas ante la inminente implantación de la evaluación de destrezas orales en el examen de lengua extranjera de la futura P.A.U.", in *Didáctica (Lengua y Literatura),* Publicaciones Universidad Complutense, 23, 11-14, http://revistas.ucm.es/index.php/DIDA/article/view/36320

MCNAMARA, Tim (1996): *Measuring second language performance*, London, Longman.

NORRIS, John, BROWN; James, HUDSON; Thom; BONK, William (2002): "Examinee abilities and task difficulty in task-based second language performance assessment", in *Language Testing,* 19, 395-418.

NORTON, Julie (2005): "The paired format in the Cambridge Speaking Test", in *ELT Journal*, 59, 287-297.

OCKEY, Gary (2009): "The effects of group members' personalities on a test taker's L2 group oral discussion test scores", in *Language Testing*, 26, 161-186.

O'SULLIVAN, Barry (2002): "Learner acquaintanceship and oral proficiency test pair-task performance", in *Language Testing*, 19, 277-295.

PEÑATE, Marcos; BAZO, Plácido (2007): "El inglés en primaria: un caso de evaluación institucional", in *Investigación en Didáctica de la Lengua y la Literatura: perspectivas y orientaciones,* MENDOZA, A.; DE AMO, J.M.; RUÍZ, M.; GALERA, F. (Eds.), Universidad de Barcelona, 311-334.

PEÑATE, Marcos; BAZO, Plácido (2014): "Using relation tables to improve validity in external evaluation" in *Porta Linguarum,* 21, 25-35.

PÉREZ JUSTE, Ramón; GARCÍA RAMOS, José Manuel (1989): *Diagnóstico, evaluación y toma de decisiones*, Madrid, Ediciones Rialp.

PLOUGH, India; BOGART, Pamela (2008): "Perceptions of examiner behavior module power relations in oral performance testing", in *Language Assessment Quarterly*, 5, 195-217.

ROBINSON, Peter (2001): "Task complexity, task difficulty and task production", in *Applied Linguistics*, 22, 27-57.

ROSS, Steven; BERWICK, Richard (1992): "The Discourse of accomodation in oral proficiency interviews", in *Studies in Second Language Acquisition*, 14, 159-176.

SAVILLE, Nick; HARGREAVES, Peter (1999): "Assessing speaking in the revised FCE", in *ELT Journal*, 53, 42-51.

STOYNOFF, Stephen (2012): "Looking backward and forward at classroom-based language assessment", in *ELT Journal,* 66, 523-532.

SUÁREZ FALCÓN, Juan Carlos (2006): "Análisis de la calidad métrica de los ítems", in *Psicometría*, BARBERO GARCÍA, M.; VILA ABAD, E. ; SUÁREZ FALCÓN, J.C. (Eds.), Madrid, Universidad Nacional de Educación a Distancia, 421-473.

SWAIN, Merrill (2001): "Examining dialogue: another approach to content specification and to validating inferences drawn from test scores", in *Language Testing*, 18, 275-302.

TAYLOR, Lynda (2001): "The paired speaking test format: recent Studies", in *University of Cambridge ESOL Examinations Research Notes*, 6, 15-17.

TAYLOR, Lynda (2005): "Washback and impact", in *ELT Journal*, 59, 154-55.

WEBB, Noreen; SHAVELSON, Richard; HAERTEL, Edward (2007): "Reliability coefficients and generalizability theory", in *Handbook of Statistics: Psychometrics,* RAO, C.R.; SINHARAY, S. (Eds.), Amsterdam, Elsevier, 81-124.

WIGGLESWORTH, Gillian; ELDER, Cathie (2010): "An investigation of the effectiveness and validity of planning time in speaking test tasks", in *Language Assessment Quarterly,* 7, 1-24.

WOOD, Manuel; BOBB, Leslie (2012): "A speaking test on the university entrance exam: how and why", in *Plurilingualism: promoting co-operation between communities, people and nations,* DIEZ, P.; PLACE, R.; FERNÁNDEZ, O. (Eds.), Universidad de Deusto, 105-126

WOOD, Manuel; PEÑATE, Marcos; BAZO, Plácido (2007): *FreconWin, Corpus Canario de Inglés Oral,* Gran Canaria, Consejería de Educación, Cultura y Deportes del Gobierno de Canarias.

YOUNG, Richard; MILANOVIC, Michael (1992): "Discourse variation in oral proficiency interviews", in *Studies in Second Language Acquisition*, 14, 403-424.

6. APPENDIX

6.1.Test 1

Rater – 1 student (one-to-one format)

Resource: a picture

Questions directly related to the picture

1) What is this group of people doing?
2) Why are they doing it in the street?
3) What is each of them doing?
4) Do you think they are close friends? How do you know?
5) Which of the two titles A or B do you prefer? Why?
   (A: 'Friends for ever', B: 'Night Street Party')

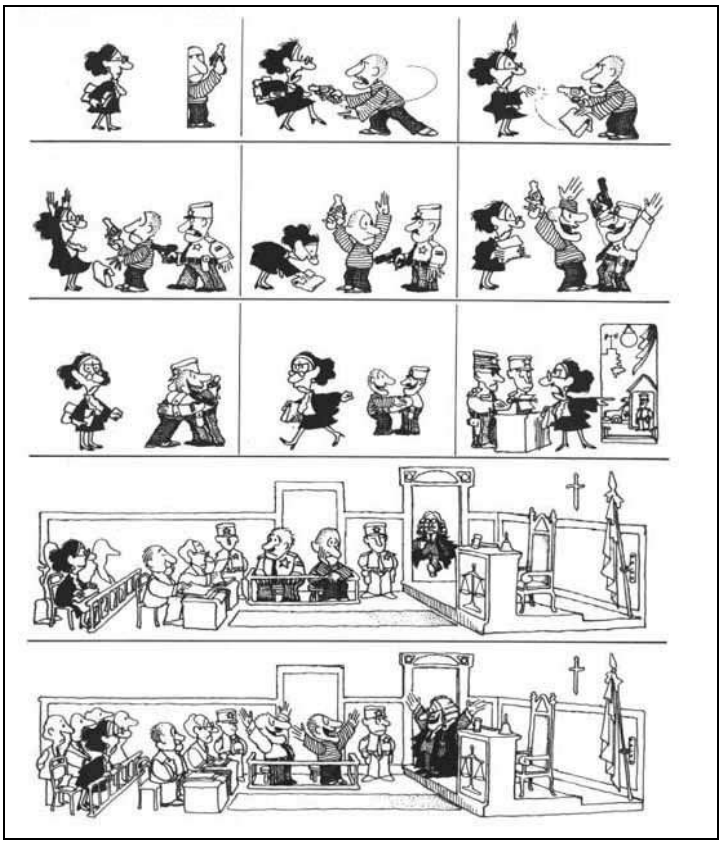
Questions indirectly related to the picture

6) What plans do you think they have made before the meeting?
7) What places are there in your area for teenagers to meet?
8) What problems does this drinking outside produce?
9) What would you do if you lived in one of these drinking places?
10) What alternatives would you suggest?

Personal related questions about the topic depicted in the picture

11) Where do you meet your friends at weekends?
12) What do you talk about during your meetings?
13) What do your parents think about the 'botellón'?
14) Would you prefer to go to a disco or a pub? Why?
15) Do you like dancing? (No —>) What do you do instead?
    (Yes—>) How many different dances do you know?

6.2. Test 2
Rater – 1 student (one-to-one format)
Resource: a comic strip



Questions directly related to the comic strip

1) Who are the main characters in the story?
2) What happens in strip one?
3) What happens in strip two?
4) What happens in strip three?
5) What happens in strips four and five?

Questions indirectly related to the comic strip

6) What lesson can you learn from the story? [About friendship, I mean] (What does the story want to say?)
7) What did the lady expect when she sat at the court before the judge came in?
8) What did the lady think when she saw the judge greeting the other two?

9) What do you think will be the end of the story?
10) Can this story be true? Why?

Personal related questions about the topic presented in the comic strip

11) Are comics only for children? Why?
12) What do you prefer a comic or a book? Why?
13) In what ways is the story told in a book different from the story told in a film?
14) What is the best book you have read? Why?
15) What is the best film you have seen? Why?

6.3. Test 3

Rater - 2 students (paired format)

Student 1

Ask your partner:

1) Con qué frecuencia suele ir de compras. (*How often do you go shopping?)*
2) Cuándo fue de compras por última vez y qué compró. (*When did you last go shopping? What did you buy?)*
3) Qué compraría si tuviera dinero para ello. (*What would you buy if you had plenty of money for it?)*
4) Cuáles son las cosas que suele comprar cuando va de compras y dónde. (*What things do you usually buy when you go shopping? Where?)*
5) Cómo disfruta / qué hace en su tiempo libre. (*What do you do in your free time? How do you enjoy your free time?*

Student 2

Ask your partner:

6) De qué trataba la última película que vio. (*What was the last film you saw about? The last film you saw, what was it about?)*
7) Que le gusta más, ver una película o leer un libro y por qué. (*What do you prefer, watching a film or reading a book? Why?*
8) La descripción de su mejor amigo. [Cómo es su mejor amigo] (*Can you describe your best friend?)*
9) Qué deporte practica en verano y dónde. (*What sport do you practise in summertime? Where?)*
10) Qué tipo de trabajo le gustaría hacer en el futuro. 8 *What kind of job would you like to have in the future?*

Rater asks the following personal related questions to both students

11) Do you pay much attention to fashion and your look?
12) What do your parents or friends tell you about your style?
13) Do you usually go to the cinema? What kind of films are you interested in?
14) Which film have you enjoyed the most lately?
15) Where do you normally go at weekends? What do you usually do there?