http://dx.doi.org/10.15446/profile.v19n_sup1.68447

Analysing Test-Takers' Views on a Computer-Based Speaking Test

Análisis de las opiniones de los candidatos sobre un examen oral a través de ordenador

Marian Amengual-Pizarro[*]
Universidad de las Islas Baleares, Palma de Mallorca, Spain

Jesús García-Laborda[**]
Universidad de Alcalá, Alcalá de Henares, Spain

This study examines test-takers' views on a computer-delivered speaking test in order to investigate the aspects they consider most relevant in technology-based oral assessment, and to explore the main advantages and disadvantages computer-based tests may offer as compared to face-to-face speaking tests. A small-scale open questionnaire was administered to 80 test-takers who took the APTIS speaking test at the Universidad de Alcalá in April 2016. Results reveal that examinees believe computer-based tests provide a valid measure of oral competence in English and are considered to be an adequate method for the assessment of speaking. Interestingly, the data suggest that personal characteristics of test-takers seem to play a key role in deciding upon the most suitable and reliable delivery mode.

Key words: Computer-based language testing, oral assessment, second language testing.

Este estudio analiza la opinión de los candidatos sobre un examen oral con ordenador para averiguar los aspectos que consideran más relevantes en la evaluación oral a través de las nuevas tecnologías y explorar las principales ventajas y desventajas de este tipo de pruebas comparadas con pruebas orales con evaluadores humanos. Se distribuyó un pequeño cuestionario a 80 candidatos que realizaron el examen oral APTIS en la Universidad de Alcalá en abril de 2016. Los resultados revelan que los candidatos consideran que las pruebas orales con ordenador son válidas y adecuadas para la evaluación de la competencia oral. Curiosamente, los datos demuestran que las características personales de los candidatos juegan un papel primordial en la elección del método de evaluación más idóneo.

Palabras clave: evaluación de lenguas a través del ordenador, evaluación de la destreza oral, evaluación de segundas lenguas.

[*] E-mail: marian.amengual@uib.es

[**] E-mail: jesus.garcialaborda@uah.es

How to cite this article (APA, 6th ed.): Amengual-Pizarro, M., & García-Laborda, J. (2017). Analysing test-takers' views on a computer-based speaking test. *Profile: Issues in Teachers' Professional Development, 19*(Suppl. 1), 23-38. http://dx.doi.org/10.15446/profile.v19n_sup1.68447.

This article was received on September 14, 2017, and accepted on November 10, 2017.

Introduction

The status of English as a global language and the new demands brought about by the Bologna Declaration (The European Higher Education Area, 1999) have led many students to take different standard English tests in order to evidence their mastery of the English language and their ability to communicate in English. Within this context, the assessment of oral skills and the development of oral language tests have received renewed interest. At the same time, over the past two decades, the use of information and communications technology (ICT) has become increasingly predominant, revolutionising the way languages are learnt, transforming educational settings, and creating new learning scenarios (Chapelle & Voss, 2016; García-Laborda, 2007; Harb, Abu Bakar, & Krish, 2014; A. C. K. Lee, 2003) and ways of learning (García-Laborda, Magal Royo, & Bakieva, 2016).

Computer technology has been especially productive in the area of language testing. As Davidson and Coombe (2012) point out, in this new era of communications technology computerised testing cannot be ignored. In fact, communications technology can provide a promising approach to test administration and delivery (García-Laborda & Martín-Monje, 2013; Zechner & Xi, 2008; Zhou, 2015). This is particularly true with regard to the assessment of students' oral production since speaking skills are commonly believed to be the most difficult and complex language abilities to test, mainly due to their specific nature (Luoma, 2004; Underhill, 1987) but also to other practicality issues such as the long time required for their evaluation, especially in high-stakes contexts (García-Laborda, 2007; Kenyon & Malone, 2010; Malabonga, Kenyon, & Carpenter, 2005; Roca-Varela & Palacios, 2013). Thus, although many language learners regard speaking as the most essential skill to be mastered (Nazara, 2011), its assessment has often been neglected in many L2 teaching and testing contexts (Amengual-Pizarro, 2009; Lewis, 2011).

As demand for oral language tests continue to grow, the integration of computer technology in the context of L2 oral assessment is gradually gaining global recognition and attention among researchers (Bulut & Kan, 2012; Zechner & Xi, 2008; Zhan & Wan, 2016). According to Galaczi (2010), the growing use of computer-based oral assessment "is largely influenced by the increased need for oral proficiency testing and the necessity to provide speaking tests that can be delivered quickly and efficiently whilst maintaining high-quality" (p. 29). The potential advantages of computer-based assessment include: higher reliability due to standardisation of test prompts and delivery, increased practicality (i.e., cost and time effective tests), faster reporting of scores, and provision of immediate feedback, among others (Araújo, 2010; García-Laborda, 2007). However, numerous concerns have also been raised over the validity of such tests (Chapelle & Douglas, 2006; Jeong et al., 2011; Zhou, 2015). Thus, computer-mediated tests are thought to limit the range of task types elicited as well as to narrow down the test construct due to the lack of an interactional component (i.e., absence of interlocutor). Indeed, the more individual view of competence highlighted in computer-delivered tests of oral proficiency seems to contradict the social oriented view of communicative performance as a jointly constructed event involving interaction between individuals (Bachman & Palmer, 1996; Chalhoub-Deville, 2003; Kramsch, 1986; McNamara, 1997). As Douglas and Hegelheimer (2007) explain, computer-based tests cannot currently capture the complexity of natural language use. Furthermore, this focus on individual competence rather than on interactional competence (Kramsch, 1986; May, 2009) may have a negative influence or washback effect on current communicative language teaching practices (Amengual-Pizarro, 2009; Green, 2013; May, 2009).

Nevertheless, some authors strongly advocate for the need to integrate computer technology in educational settings. Furthermore, Chapelle and Douglas (2006) claim that "communicative language ability needs to be conceived in view of the joint role that language and technology play in the process of communication" (p. 108). In this regard, numerous researchers point out that the advantages of using computer-based technology can outweigh some of its limitations (García-Laborda, 2007; García-Laborda, Magal Royo, Litzler, & Giménez López, 2014) as well as the positive attitude of many test takers (Litzler & García-Laborda, 2016). Under this perspective, computer language testing is presented as a feasible alternative to other traditional methods of testing oral skills such as face-to-face assessment since it clearly facilitates test administration and delivery by reducing testing time and costs (Araújo, 2010; Bulut & Kan, 2012; García-Laborda, 2007; Qian, 2009). Furthermore, various research studies provide evidence of score equivalence between the two types of delivery modes (computer-based tests vs. face-to-face tests) on the testing of oral skills (Shohamy, 1994; Zhou, 2015). Thus, numerous examination boards have started to develop computer-based oral assessment: The Computerised Oral Proficiency Instrument (COPI), the Pearson Test of English (PTE) (Pearson, 2009a), the Versant tests (Pearson, 2009b), the TOEFL iBT speaking test (Zechner, Higgins, & Williamson, 2009), the APTIS speaking test (O'Sullivan & Weir, 2011), and so on. Bernstein, Van Moere, and Cheng (2010) also support the validity of some fully automated spoken language tests by establishing a construct definition for these types of tests (see Lamy, 2004) and providing concurrent data relating automated test scores to communicative tests. These latter authors suggest that automated test scores can be used in a complementary way with other forms of assessment in decision making. In the same vein, Galaczi (2010) explains that computer-based tests can effectively be used to supplement other more traditional speaking language tests.

Taking these findings as a basis, the main aim of this paper is to examine candidates' views on a computer-based speaking test (the APTIS speaking test) in order to gain a better insight about the advantages and disadvantages computer-mediated tests may offer as compared to more traditional face-to-face speaking tests (i.e., oral interviews), and to explore the aspects test-takers consider most relevant in technologically enhanced oral language tests.

The APTIS Test

This paper examines test-takers' opinions on the implementation of APTIS, a computer-based test of general English proficiency developed by the British Council (see O'Sullivan, 2012; O'Sullivan & Weir, 2011). APTIS intends to offer an alternative to high-stakes certificated tests designed for a population over age 15 and it comprises five main components: core (grammar and vocabulary), listening, reading, writing, and speaking. Although APTIS can be administered in more traditional ways such as pen-and-paper, it is usually taken via computer.

In order to report APTIS test results, a numerical scale (0-50) is used following the Common European Framework of Reference for Languages (CEFR) to test language abilities across the A1-B2 range. Test results are usually reported within 48 hours.

The APTIS speaking test takes around 12 minutes to complete and is divided into four sections (see Table 1). Responses are recorded and marked holistically on a six-point (Tasks 1 to 3) and a seven-point scale (Task 4) by a certified human examiner.

Table 1. Components of the APTIS Speaking Test (Descriptive Statistics)

Section	Technique	No. of Items & Time
1	Personal information	Three questions, 30 seconds each question
2	Photograph description and comparison	Different number of questions, 45 seconds for each question
3	Picture comparison	Two questions, 45 seconds for each question
4	Questions based on an image (single topic)	Three questions, 2 minutes (1-minute preparation time)

As can be seen, Part 1 of the examination involves three questions on various personal topics in order to relax students and get them used to talking on a computer. Candidates are given 30 seconds to answer each question (three questions). Part 2 requires test-takers to describe and compare different pictures (i.e., picture description). Questions in this section may range in difficulty. Candidates are allowed 45 seconds to answer each question. Part 3 consists of the comparison of two pictures (i.e., describe, compare, and speculate). Candidates are asked two questions. The last question usually involves imaginary situations or speculation. Again 45 seconds are allowed for each question in this section. Finally, Part 4 consists of three questions on a single topic (e.g., personal abstract ideas). Test-takers are given one minute to prepare their response and are allowed to make brief notes to structure their answers. They are expected to talk for two minutes.

Research Questions

As previously noted, the main purpose of this paper is to study test-takers' views on the use of a computer delivery oral test (the APTIS speaking test), and to explore the main differences between computer-delivery vs. face-to-face mode on the assessment of speaking. More specifically, the following aspects were addressed:

1. Use of preparation material for the computer-based speaking test.
2. Assessment of oral skills via computer.
3. Usefulness of note-taking and exam simulation prior to the official computer-based test of speaking.
4. Degree of complexity of the computer programme.
5. Usefulness of self-evaluation sessions prior to the actual computer-based test.
6. Main differences between the computer-based test (i.e., APTIS speaking test) and the face-to-face test (i.e., interview with an examiner)

Method

Participants

A total of 80 students at the Universdad de Alcalá (Madrid, Spain) took part in this study. As regards gender distribution, the majority of participants were females (85%, *n* = 68), and 15% (*n* = 12) males. Most of the participants ranged in age from 18 to 22 years (65.2%); 17% were from 23 to 25 years of age, and 9.4% were over 25 years old. The remaining 8.4% of the participants did not provide an answer to this question.

Data Collection Instrument

A small-scale open questionnaire (see Appendix) was distributed to the participants by computer delivery mode in mid-April 2016 in order to capture their opinion on the APTIS speaking test, and to examine the main differences between computer-assisted vs. face-to-face speaking assessment. Participants were given two days to enter their answers on a computer and send them back to the researchers after having taken the official APTIS speaking test. All participants had been previously interviewed by the researchers in early February 2016 to determine their levels of spoken English. The tasks included in the personal interviews were similar to the ones featured in the APTIS speaking test, namely: some warm-up questions on a personal topic, a photograph description, a comparison of two photographs, and a discussion of concrete and more abstract topics.

The questionnaire contained 17 questions related to the main following aspects: (a) use of material to prepare for the computer-based speaking test (Items 1 and 2); (b) assessment of oral skills via computer (Items 3 and 4); (c) usefulness of note-taking and exam simulation prior to the official computer-administered oral test (Items 5 to 10); (d) degree of complexity of the computer programme (Items 11 and 12); (e) usefulness of self-evaluation sessions prior the actual computer-based test (Items 13 to 15), and (f) main differences between the computer-based test and the face-to-face test (Items 16 and 17).

Participants were asked to rate the first five main aspects (Items 1 to 15) on a 1-4 Likert scale ranging from 1 (totally disagree) to 4 (totally agree). Additional qualitative comments could also be provided by respondents in order to elaborate their answers on some questions and help researchers to get a better insight of the data provided. The remaining two questions of the questionnaire (Items 16 and 17) were explored by means of two open-ended questions. The questionnaire was administered in Spanish since this is the communication language of the participants. The reliability of the questionnaire had a Cronbach's alpha of 0.769, which indicates a relatively high level of internal consistency. Quantitative results were analysed using the Statistical Package for the Social Sciences (SPSS) 21.0 programme.

Results

Descriptive statistics are presented first followed by the qualitative analyses of the examinees' comments to the two open-ended questions included in the questionnaire (Items 16 and 17).

Quantitative Results

Use of Material to Prepare for the Computer-Based Speaking Test

The first section of the questionnaire (Items 1 and 2) attempted to examine tests-takers' opinion on the exam-related materials for the APTIS speaking test provided by researchers. The mean scores and standard deviations were calculated for each item (Table 2).

As can be seen, the mean scores of the two items in Table 2 are above 2.5 on a 4-point scale which indicates that both aspects were regarded as relevant in order to obtain good test results. Thus, respondents admitted making a great use of the support material provided by researchers (Item 1; $\bar{x} = 2.89$) and considered this guidance material to be helpful (Item 2; $\bar{x} = 2.61$) since it assisted them in becoming familiar with the format of the test and its level of difficulty. Overall, the data suggest that all the candidates made use of test-related material to do their best on the test and succeed in the examination.

Assessment of Oral Skills Via Computer

Items 3 and 4 in the questionnaire intended to determine participants' opinion on the validity and suitability of computer-based tests to assess speaking.

The results presented in Table 3 indicate that participants believe computerised testing can adequately measure their oral skills and, therefore, it is considered to be both a valid (i.e., face validity) (Item 4: "Computerised testing measures my spoken ability in English effectively"; $\bar{x} = 2.29$), as well as a suitable method for the assessment of speaking ability in English (Item 3): "The computer is an appropriate method for

Table 2. Use of Preparation Material (Descriptive Statistics)

	N	Minimum	Maximum	Mean	*SD*
1. Use of support material	76*	1	4	2.89	0.930
2. Usefulness of support material	76*	1	4	2.61	0.943

*4 missing cases.

Table 3. Use Material (Descriptive Statistics)

	N	Minimum	Maximum	Mean	*SD*
4. Computer-test validity	78*	1	4	2.29	0.870
3. Computer-test suitability	80	1	4	2.28	1.031

*2 missing cases.

the APTIS speaking test", $\bar{x}$ = 2.28). Indeed, results also show a reasonable significant correlation (r = 0.526) at p = 0.01 between these two variables. However, the higher standard deviation on Item 4 (SD = 1.031) indicates a major dispersion or variation of the data around the mean. That is, there seems to be a higher consensus among participants on the validity of the test (i.e., face validity, Item 4) rather than on the adequacy of using a computer-administered test to assess oral skills (Item 3). The optional qualitative comments provided by respondents point to possible reasons that could help us understand the main discrepancies regarding this issue. Thus, respondents who favoured computerised testing highlighted the potential advantage of having their performances recorded since this was thought to help examiners to listen to the test recordings as many times as necessary before they decided on their final score. Some other examinees also reported performing better before a computer since they felt less nervous than in more traditional face-to-face speaking test situations. On the contrary, many test-takers consider the absence of an interlocutor to interact with, and receive some feed-back from, as a negative aspect which may hinder their performance and affect their test scores in a detrimental way. In any event, the mean values of Items 4 and 3 are above two points on a 4-point scale which indicate participants' overall positive views on both aspects.

Usefulness of Note-Taking and Exam Simulation Prior to the Official Computer-Based Test of Speaking

As far as the usefulness of note-taking and exam simulation prior to the official APTIS speaking test is concerned, the data (Table 4) reveal that both aspects, along with the training sessions provided by researchers, were highly regarded by participants.

Results in Table 4 have been arranged in descending order of importance so as to facilitate comprehension.

Among the most useful tasks, examinees ranked the following in order of importance: Taking notes prior to the recorded simulation (Item 6; $\bar{x}$ = 3.34), taking a mock test before sitting for the actual test itself (Item 5; $\bar{x}$ = 3.23), and having a training session prior to the official examination in order to familiarise them with the testing procedure and help them obtain better test results (Item 10; $\bar{x}$ = 2.95).

As shown in Table 4, taking notes during the official test (Item 7; $\bar{x}$ = 2.90) was felt to favour exam results to a lesser extent than taking notes during the mock test (Item 6; $\bar{x}$ = 3.34). The qualitative comments provided by participants in this regard point to the tight time frame set for taking notes during the official computer-based test. This is an important aspect to bear in mind since research suggests that test-takers may experience a negative affect due to inadequate or insufficient planning time (Malabonga et al., 2005).

Table 4. Use of Notes and Practice Exam (Descriptive Statistics)

	N	Minimum	Maximum	Mean	*SD*
6. Usefulness of note-taking prior to mock test	80	1	4	3.34	0.973
5. Usefulness of mock test	77*	1	4	3.23	0.887
10. Usefulness of training sessions	77*	1	4	2.95	0.759
7. Usefulness of note- taking during APTIS test	80	1	4	2.90	1.051
8. Reading notes during mock test	79*	1	4	2.26	0.999
9. Reading notes during aptis test	79*	1	4	2.11	0.920

*Items containing missing cases.

Table 5. Evaluation of Computer Programme (Descriptive Statistics)

	N	Minimum	Maximum	Mean	*SD*
12. Intuitive software application	80	2	4	3.53	0.596
11. User-friendly software application	80	2	4	3.49	0.638

Clearly, the pressure associated with each testing situation (mock test vs. official test) may play a key role in interpreting examinees' answers to that question. Interestingly, scores for the APTIS speaking test were found to be higher than the scores for the mock test, although no statistically significant differences were found between both examinations. On the whole, test-takers agreed that note-taking was useful to help them structure their ideas, recall some useful expressions, and avoid improvisations.

As can be observed, the lowest two ranking items are related to the use of participants' test notes both during the mock exam (Item 8; $\bar{x}$ = 2.26) and the actual APTIS test (Item 9; $\bar{x}$ = 2.11). Thus, participants admitted not speaking fluently and reading from their notes mainly during the practice test. This might be due to personal factors such as nervousness or a lack of previous experience in computerised assessment before the official test took place. However, it is believed that some attention should be paid to this aspect in order to prevent test-takers from writing out their complete answers and memorising words and expressions that might affect natural target language use.

Degree of Complexity of the Computer Programme

In order to examine the degree of complexity of the test computer programme, examinees were asked to rank two items (Items 11 and 12). The findings are presented in Table 5.

As can be observed, participants clearly favour the logistical advantages provided by the computerised format of the test. In fact, these are the two highest ranking items of the questionnaire. Furthermore, none of the answers provided by respondents registered an extreme negative value (minimum value = 1). Test-takers considered that the application presented no operational difficulties and felt the software was very intuitive (Item 12; $\bar{x}$ = 3.53) and reasonable to handle (Item 11; $\bar{x}$ = 3.49). The technological simplicity of the APTIS test may seem to be an advantage for most test-takers (see Kenyon & Malabonga, 2001) who appeared to feel comfortable with the management of the new software. This is an important aspect to be taken into account since various research findings suggest that computer familiarity and other features of the computerised context (e.g., computer anxiety) may affect candidates' performance (Chapelle, 2001; Clariana & Wallace, 2002; Colwell, 2013; J. A. Lee, 1986; Norris, 2001; Taylor, Kirsch, Jamieson, & Eignor, 1999). As Litzler and García-Laborda (2016) point out: "students need to understand how the software works in addition to knowing the content of the exam, which can be difficult without previous training" (p. 107). Otherwise, the technological mediation of the testing process can prevent test-takers from demonstrating their real proficiency level in spoken English.

Usefulness of Self-Evaluation Sessions Prior to the Actual Computer-Based Test

Participants were also required to assess the suitability of the self-assessment sessions carried out by researchers to help them determine their actual level of spoken English. These sessions also aimed at encouraging self-reflection and promoting the effective implementation of different test strategies to ensure the achievement of the best results on the computer-based test. As shown in Table 6, the data reveal that test-takers seem to hold very positive views on the self-evaluation sessions conducted by researchers, especially as far as the acquisition of learning strategies (Item 15; $\bar{x} = 3.19$) is concerned. This stresses the importance of developing metacognitive strategies to raise examinees' awareness of their strengths and weaknesses so as to be able to improve their test behaviour and do their best on the test.

Qualitative Results

Differences Between the Computer-Based Test (APTIS Test) and the Face-to-Face Test

A comparison of the test-takers' comments was made to address research question number 6 (Items 16 and 17). Question 16 queried students about the main differences between the APTIS speaking test and the face-to-face speaking test. Responses to this open-ended question reveal advantages and disadvantages of both types of delivery modes.

Here is a sampling of the more negative aspects related to the APTIS speaking test:

> I prefer to talk to a human person rather to a computer since a person can inspire confidence. Computers are cold and they do not give you any feedback (they do not look or smile at you…). Computers cannot help you either. If you do not know what to say, they are not going to try to help you. (s7)
>
> The main difference between the two delivery modes is that there seems to be a stricter control of time in the implementation of the computer-based test which turns out to be very stressing. (s37)
>
> In computer-based testing there is no interaction. I prefer human-delivered speaking tests because examiners can offer you some help in case you get stuck or can give you some clues on how to interpret certain words or images on certain occasions. (s50)

However, other test-takers favoured the computerised format of the test over the face-to-face speaking test:

> The computer-based test is more dynamic. You can organise yourself better and I think it is much more efficient. (s2)
>
> Computer-based testing is much more comfortable and less stressful than face-to-face speaking tests. I am an introvert person and talking to a computer makes me feel less embarrassed because nobody can laugh at me. (s28)
>
> In human-delivered speaking tests there is more tension, you can see the examiner looking at you all the time as well as his/her expressions, which can be very distracting and stressful. (s48)

To sum up, the main advantages of face-to-face assessment appear to be related to interaction, authenticity (i.e., real communication), and provision of helpful feedback which seems to be lacking in computerised testing. This is in line with research findings that suggest that this latter kind of delivery mode may be found "de-humanizing" by examinees (Kenyon & Malabonga, 2001). In computer-based assessment, the strict control of time also seems to be of concern to some test-takers. In fact, numerous participants found the timer on the screen very stressful. On the contrary, some other examinees believed computer-delivered tests were very convenient, dynamic, and effective. The absence of a human examiner looking at candidates and taking notes was also considered a positive aspect by these latter participants.

Finally, the last item of the questionnaire (Item 17) asked test-takers to express their opinions on the form of assessment (computer-assisted testing vs. face-to-face assessment) they believed was more efficient to evaluate their oral skills. Similar to previous research findings

Table 6. Self-Evaluation Sessions (Descriptive Statistics)

	N	Minimum	Maximum	Mean	*SD*
15. Use of strategies after self-assessment session	76*	1	4	3.19	0.824
13. Suitability of self-assessment session	77*	1	4	2.85	0.780
14. Self-assessment of competence in TL	74*	1	4	2.61	0.784

*Items containing missing cases.

(Qian, 2009; Shohamy, 1994), face-to-face speaking tests drew the most positive results on this aspect. Thus, 75% of test-takers favoured face-to-face speaking tests over technology-mediated speaking tests since, as previously anticipated, oral communication is mainly regarded as a human phenomenon (Kenyon & Malabonga, 2001). The main reasons provided by candidates in this respect were mostly related to the interactional nature of conversation (i.e., use of real or "authentic" language, importance of gestures and body language, and provision of feedback), which could not be appropriately captured in technology-mediated assessment (Douglas & Hegelheimer, 2007; Kenyon & Malabonga, 2001). These findings point to the importance attached by candidates to interpersonal cues and to the negotiation of meaning between interlocutors to interact and reach communicative goals (Ockey, 2009; Qian, 2009).

A minority of test-takers also commented that they felt disadvantaged by talking into a computer since computer-based tests could present some logistical or technical problems such as sound or audio problems. Other examinees seemed to believe that the traditional testing format is just more reliable and efficient (see Colwell, 2013). Some of the test-takers' comments regarding this issue are the following:

> I think face-to-face speaking tests can better assess spoken competence because examiners can see the way you talk or the body language you use which is a key element in real communication, and cannot be captured by a computer. (s17)
>
> I think the presence of an examiner is very helpful because they can see your gestures, the way you talk and they can help you by asking some questions in case you do not know what to say. A computer, however, cuts you off abruptly after 45 seconds and you are not given the opportunity to show your true oral skills. (s27)
>
> I think face-to-face speaking tests are more effective because there are always some problems with the microphones and the recordings sometimes cannot be heard appropriately. (s6)

Interestingly, the main reason put forward by those candidates who favoured computer-administered tests (16.3% of the respondents) was clearly related to personal characteristics of test-takers such as introversion or embarrassment. Thus, many test-takers described themselves as shy or introverted and pointed out they felt more relaxed before a computer without the presence of an examiner. This finding is consistent with previous research which suggests that candidates' personal characteristics such as level of extroversion or introversion can affect their test scores in oral assessments (Kang, 2008; Nakatsuhara, 2010; O'Sullivan, 2000; Ockey, 2009; Underhill, 1987). Other test-takers were also positive about the use of computerised tests since this type of assessment was thought to increase rater reliability. In fact, some respondents pointed out that the use of computers could help prevent raters from being influenced by candidates' personal characteristics, as was often believed to be the case in human-scored tests (see also Lumley & McNamara, 1995; Stemler, 2004). Some illustrative comments related to the potential benefits of computer-based testing are as follow:

> I think computer-based speaking tests are more effective because I reckon I am a very shy person and I express myself worse before an examiner rather than in front of a computer. Therefore, I get worse results in face-to-face assessment. (s31)
>
> I feel more comfortable doing computer-based tests because I do not have the feeling of being constantly observed by an examiner. (s34)
>
> I think computerised testing is more reliable. Examiners cannot see our faces and they can therefore be more objective. (s79)

According to various research findings, the introduction of new technology may be seen to add further difficulties to the test causing unnecessary stress and uncertainty to candidates (Bartram, 2006; Litzler & García-Laborda, 2016; Saadé & Kira, 2007). Interestingly, some test-takers commented that apprehension toward computer-based testing was greatly reduced after having taken the APTIS speaking test on the computer (see Bernstein et al., 2010; Zhou, 2015). The following comment is an illustrative example:

> I used to prefer face to-face tests but, surprisingly, after sitting for the APTIS test I think I feel now quite comfortable taking computer-based tests. (S78)

Finally, a small number of participants (8.8%) stressed the advantages of both delivery modes, highlighting the potential applicability of computer-administered assessment in different testing contexts:

> I think both types of assessment are effective. I believe there are factors other than the type of test delivery mode that may have a greater influence on scores, such as the selection of questions used to demonstrate your speaking abilities. (S29)
>
> In my view, it depends on the context. If the purpose of the test is to assess the real spoken competence of the student in a physical context (i.e., conversation in the street, at the office, etc.) then, I think it is better the face-to-face test. But if you want to assess oral skills on an audio-visual context (i.e., Skype, etc.) then, I believe it is better to use technology-based tests. (S17)

Indeed, this latter comment (S17) points to the importance of construct definition for computer-administered oral proficiency tests. As some authors suggest, test validity is an aspect necessarily linked to the use of scores. Likewise, Bernstein et al. (2010) explain: "Validity can only be established with reference to test score use and decisions made on the basis of test scores, and not merely on the basis of consistently measuring test-takers according to a defined construct" (p. 372).

In short, the major advantage of face-to-face tests seems to be related to the possibility of human interaction which enhances authenticity and reflects the communicative nature of language use. Participants also appreciate the possibility of having some feedback from the examiners which might encourage some candidates to feel at ease and to get to talk in case they do not know what to say. These are the main reasons why the majority of test-takers believe face-to-face speaking tests allow a more effective evaluation of their actual oral competence in English. However, the presence of an interlocutor can also negatively affect tests-takers' behaviour and add further pressure, especially to more introverted candidates. Furthermore, for those latter participants, computerised testing is felt to produce more reliable results since examiners cannot be influenced by candidates' personal characteristics.

Conclusion

The findings of this study reveal that, despite the difficulty of capturing human oral language interaction (Douglas & Hegelheimer, 2007; Kenyon & Malabonga, 2001), computer-administered tests are thought to provide a valid measure of oral competence in English (i.e., face validity) and to be an appropriate method for the assessment of oral skills. More specifically, the results show that on the whole participants hold positive views on the APTIS speaking test and consider the test application to be very convenient, intuitive, and user-friendly. The data also reveal that candidates have very favourable opinions of the support material used by researchers to familiarise them with the test format, content, and level of difficulty of the examination. Furthermore, they believe that the training sessions for self-reflection and development of learning strategies proved to be very useful to obtain good results on the test. These are important aspects to bear in mind in order to reduce the potential negative influence related to the technological mediation of the testing process (Bartram, 2006; Chapelle, 2001; Norris, 2001; Saade & Kira, 2007). Admittedly, test-takers clearly favour face-to-face tests over computer-administered tests for the assessment of oral ability due to the intrinsically social and interactional nature of speaking skills (McNamara, 1997), which do not appear to be effectively elicited in computerised formats (Araújo, 2010; Kenyon & Malabonga, 2001). Interestingly, the findings suggest that personal characteristics of test-takers such as introversion may play a key role in deciding upon the most suitable delivery mode for the assessment of oral language skills since introverted candidates reported feeling less anxious and much more comfortable without the presence of an interlocutor.

The majority of these latter participants also believe computerised-test scores tend to be more reliable due to the fact that examiners cannot be influenced by test-takers' personal characteristics.

On the whole, these are encouraging results since they seem to confirm that technology-based tests can be used as an efficient complement to face-to-face assessment in order to evaluate speech production. As Galaczi (2010) reminds us, a key concept in language testing is "fitness for purpose": "Tests are not just valid, they are valid *for* a specific purpose, and as such, different test formats have different applicability for different contexts, age groups, proficiency levels, and score-user requirements" (p. 26). In the same vein, numerous researchers highlight the importance of establishing construct validity with reference to the inferences and decisions made on the basis of test scores (Bernstein et al., 2010; Galaczi, 2010; Xi, 2010). As Bernstein et al. (2010) point out, computer test scores provide one piece of evidence about a candidate's performance but should not be necessarily used as the only basis of decision-making. Therefore, it is believed that both types of methods should be seen as complementary rather than as competing alternatives. Indeed, computer-based tests may offer a promising methodology to assist face-to-face speaking tests, contributing to decision-making in multiple educational and occupational contexts.

References

Amengual-Pizarro, M. (2009). Does the English test in the Spanish university entrance examination influence the teaching of English? *English Studies*, *90*(5), 585-598. https://doi.org/10.1080/00138380903181031.

Araújo. L. (Ed.). (2010). *Computer-based assessment (CBA) of foreign language speaking skills*. Luxembourg, LU: Publications Office of the European Union. Retrieved from https://www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/Volume_European_Commission_2010.pdf.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bartram, D. (2006). The internationalization of testing and new models of test delivery on the internet. *International Journal of Testing*, *6*(2), 121-131. https://doi.org/10.1207/s15327574ijt0602_2.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355-377. https://doi.org/10.1177/0265532210364404.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate students in Turkey. *Eurasian Journal of Educational Research*, *49*, 61-80.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*, 369-383. https://doi.org/10.1191/0265532203lt264oa.

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139524681.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511733116.

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in language learning and technology. *Language Learning & Technology*, *20*(2), 116-128. Retrieved from http://llt.msu.edu/issues/june2016/chapellevoss.pdf.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602. https://doi.org/10.1111/1467-8535.00294.

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing and the common core. *Journal of Education and Training Studies*, *1*(2), 50-60. https://doi.org/10.11114/jets.v1i2.101.

Davidson, P., & Coombe, C. (2012). Computerized language assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second* *language assessment* (pp. 267-273). Cambridge, UK: Cambridge University Press.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27*, 115-132. https://doi.org/10.1017/S0267190508070062.

The European Higher Education Area. (1999). *The Bologna declaration of 19 June 1999: Joint declaration of the European Ministers of Education*. Retrieved from http://www.magna-charta.org/resources/files/BOLOGNA_DECLARATION.pdf.

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

García-Laborda, J., Magal Royo, M. T., & Bakieva, M. (2016). Looking towards the future of language assessment: Usability of tablet PCs in language testing. *Journal of Universal Computer Science, 22*(1), 114-123.

García-Laborda, J., Magal Royo, T., Litzler, M. F., & Giménez López, J. L. (2014). Mobile phones for a university entrance examination language test in Spain. *Journal of Educational Technology & Society, 17*(2), 17-30.

García-Laborda. J., & Martín-Monje. E. (2013). Item and test construct definition for the new Spanish baccalaureate final evaluation: A proposal. *International Journal of English Studies*, *13*(2), 69-88. https://doi.org/10.6018/ijes.13.2.185921.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Computer-based assessment of foreign language speaking skills* (pp. 29-51). Luxembourg, LU: Publications Office of the European Union. Retrieved from https://www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/Volume_European_Commission_2010.pdf.

Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, *13*(2), 39-51. https://doi.org/10.6018/ijes.13.2.185891.

Harb, J., Abu Bakar, N., & Krish, P. (2014). Gender differences in attitudes towards learning oral skills using technology. *Education Information Technology, 19*(4), 805-816. https://doi.org/10.1007/s10639-013-9253-0.

Jeong, H., Hashizume, H., Sigiura, M., Sassa, Y., Yokoyama, S., Shiozaki, S., & Kawashima, R. (2011). Testing second oral language proficiency in direct and semi-direct settings: A social cognitive neuroscience perspective. *Language learning*, *61*(3), 675-699. https://doi.org/10.1111/j.1467-9922.2011.00635.x.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology, 5*(2), 60-83.

Kenyon, D. M., & Malone, M. (2010). Investigating examinee autonomy in a computerized test of oral proficiency. In L. Araújo (Ed.), *Computer-based assessment of foreign language speaking skills* (pp. 1-27). Luxembourg, LU: Publications Office of the European Union. Retrieved from https://www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/Volume_European_Commission_2010.pdf.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 181-205.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*(4), 366-372. https://doi.org/10.1111/j.1540-4781.1986.tb05291.x.

Lamy, M.-N. (2004). Oral conversations online: Redefining oral competence in synchronous environments. *ReCALL, 16*(2), 520-538. https://doi.org/10.1017/S095834400400182X.

Lee, J. A. (1986). The effects of past computer experience on computer aptitude test performance. *Educational and Psychological Measurement, 46*, 727-736. https://doi.org/10.1177/0013164486463030.

Lee, A. C. K. (2003). Undergraduate students' gender differences in IT skills and attitudes. *Journal of Computer Assisted Learning, 19*(4), 488-500. https://doi.org/10.1046/j.0266-4909.2003.00052.x.

Lewis, S. (2011). Are communication strategies teachable? *Encuentro, 20*, 46-54.

Litzler, M. F., & García-Laborda, J. (2016). Students' opinions about ubiquitous delivery of standardized English exams. *Porta Linguarum,* (Monográfico 1), 99-110.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language testing, 12*(1), 54-71. https://doi.org/10.1177/026553229501200104.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511733017.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing, 22*(1), 59-92. https://doi.org/10.1191/0265532205lt297oa.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 387-421. https://doi.org/10.1177/0265532209104668.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446-466. https://doi.org/10.1093/applin/18.4.446.

Nakatsuhara, F. (2010, April). *Interactional competence measured in group oral tests: How do test-talker characteristics, task types and group sizes affect co-constructed discourse in groups*. Paper presented at The Language Testing Research Colloquium, Cambridge, United Kingdom.

Nazara, S. (2011). Students' perception on EFL speaking skill development. *Journal of English Teaching, 1*(1), 28-42.

Norris, J. M. (2001). Concerns with computerized adaptive oral proficiency assessment: A commentary on "Comparing examinee attitudes toward computer-assisted and oral proficiency assessments" by Dorry Kenyon and Valerie Malabonga. *Language learning & Technology, 5*(2), 99-105. Retrieved from http://llt.msu.edu/vol5num2/pdf/norris.pdf.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*(3), 378-386. https://doi.org/10.1016/S0346-251X(00)00018-X.

O'Sullivan, B. (2012). *Aptis test development approach* (Aptis technical Report, ATR-1). London, UK: British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/aptis-test-dev-approach-report.pdf.

O'Sullivan, B., & Weir, C. (2011). Language testing and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13-32). Oxford, UK: Palgrave.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161-186. https://doi.org/10.1177/0265532208101005.

Pearson. (2009a). *Official guide to Pearson test of English academic*. London, UK: Author.

Pearson. (2009b). *Versant Spanish test: Test description and validation summary.* Palo Alto, US: Author.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113-125. https://doi.org/10.1080/15434300902800059.

Roca-Varela, M. L., & Palacios, I. M. (2013). How are spoken skills assessed in proficiency tests of general English as a foreign language? A preliminary survey. *International Journal of English Studies, 13*(2), 53-68. http://dx.doi.org/10.6018/ijes.13.2.185901.

Saadé, R. G., & Kira, D. (2007). Mediating the impact of technology usage on perceived ease of use by anxiety. *Computers & Education, 49*(4), 1189-1204. https://doi.org/10.1016/j.compedu.2006.01.009.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99-123. https://doi.org/10.1177/026553229401100202.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4.

Taylor, C., Kirsch, I., Jamieson, J., & Eignor, D. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *A Journal of Research in Language Studies, 49*(2), 219-274. https://doi.org/10.1111/0023-8333.00088.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques.* Cambridge, UK: Cambridge University Press.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291-300. https://doi.org/10.1177/0265532210364643.

Zechner, K., & Xi, X. (2008, June). Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the third ACL workshop on innovative use of NPL for building educational applications* (pp. 98-106). Stroudsburg, US: Association for Computational Linguistics. https://doi.org/10.3115/1631836.1631848.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883-895. https://doi.org/10.1016/j.specom.2009.04.009.

Zhan Y., & Wan, Z. H. (2016). Test takers' beliefs and experiences of a high-stakes computer-based English listening and speaking test. *RELC Journal, 47*(3), 363-376. https://doi.org/10.1177/0033688216631174.

Zhou, Y. (2015). Computer-delivered or face-to-face: effects of delivery-mode on the testing of second language speaking. *Language Testing in Asia, 5*(2). https://doi.org/10.1186/s40468-014-0012-y.

About the Authors

Marian Amengual-Pizarro holds a PhD in English linguistics and is an Associate Professor at Universidad de las Islas Baleares, Spain. She has contributed numerous articles on testing and education to various specialized refereed international journals. She is currently the editor of the journal *Revista Electrónica de Lingüística Aplicada (RAEL)*.

Jesús García-Laborda is an Associate Professor at Universidad de Alcalá, Spain. He has a Doctorate in English and another in Language Education. His research focuses on computer-based language testing, English for specific purposes, and teacher education. He has published numerous articles in these fields in specialized refereed international journals.

Appendix: APTIS Questionnaire

Please say to what extent you agree with the following statements by circling a number from 1 (completely disagree) to 4 (completely agree). Please do not leave out any of the items.

Name: **Age:** **Sex:** Male □ Female □

A. Preparation for the APTIS speaking test

1. I used the support material provided by researchers.	1 2 3 4
2. Now that I have taken the official APTIS exam, I can definitely say that the material provided by researchers really helped me.	1 2 3 4

B. Computer delivery mode

3. The computer is an appropriate method for the APTIS speaking test. Justify your answer:	1 2 3 4
4. Computerised testing measures my spoken ability in English effectively. Justify your answer:	1 2 3 4

C. Use of notes and exam simulation prior to speaking test

5. The mock exam I took prior to sitting for the official APTIS test has helped me to do well on the examination.	1 2 3 4
6. Taking notes before recording the examination during the training session helped me to perform better during the mock test. In what way did taking notes help you?	1 2 3 4
7. Taking notes during the official APTIS speaking test helped me to perform better during the actual test. In what way did taking notes help you?	1 2 3 4
8. I read my notes during the mock test (that is, I did not speak fluently during the test).	1 2 3 4
9. I read my notes during the official APTIS test when performing the oral tasks.	1 2 3 4
10. The training sessions for the APTIS speaking test helped me to obtain a good test result.	1 2 3 4
11. The APTIS speaking software application is user-friendly.	1 2 3 4
12. The APTIS speaking software application is intuitive.	1 2 3 4

D. Self-evaluation sessions	
13. The setup of the self-assessment session was adequate.	1 2 3 4
14. The self-assessment session helped me to determine my actual level of spoken English.	1 2 3 4
15. The self-assessment session helped me to develop strategies to improve my performance during the test.	1 2 3 4

Now, please answer the two following questions as honestly as possible:

16. What are the main differences between the computer-administered test and the face-to-face test (i.e., interview with an examiner)?

17. What type of assessment (computer-based test vs. face-to-face test) do you think is better to evaluate your real ability in spoken English?

Thanks for your collaboration!!