

ON THE NET

Introducing Standardized EFL/ESL Exams

Jesús García Laborda

Polytechnic University of Valencia, Spain

INTRODUCTION

In a few months students from almost every single country will be taking high-stakes tests such as the *Test of English as a Foreign Language* (TOEFL), *International English Testing System* (IELTS), *Business Language Testing Service* (BULATS), and others in order to achieve a score that allows them to pursue their studies abroad. Students who want to pursue graduate studies in the United States will be able to choose from a number of tests to prove their proficiency level, but, up to now, the most important test has been the TOEFL. The TOEFL, like other exams¹, has been progressively converted into a computer-based test (CBT), first, and into an Internet-based test (iBT) later. Most students today take a computer version of this test. The promptness of the rating system as compared to the traditional paper-and-pencil format have helped to popularize the CBT and iBT versions of TOEFL. Many of the other high-stakes tests have followed or are following this change. Probably, the second most important high-stakes exam is IELTS, whose computerized version is currently being used experimentally in several countries as a first stage of implementation.

The degree of computerization varies according to the test. It runs from an implementation of the four skills (iBT TOEFL) to only the least communicative sections (BULATS), with some exams using a combination of human and computer interaction such as the Basic English Skill Test Plus (BEST Plus). The continuous evolution of the Internet and computer devices, exam takers' better training, and the reduced costs of implementing computerized tests will eventually facilitate the transition from the current paper format to the electronic one.

This paper will present the features, and a brief comparison, of some of the most well known high-stakes exams. Since some of these tests are not well known or are not easily accessible to the general practitioner, this paper will serve as a first approach to a field that will undoubtedly change significantly in the next five years.

COMPUTER TEST DESIGN

Computer tests have been operative for at least 20 years. However, although the potential of computer assisted language testing was foreseen from the mid-eighties, it has been necessary to develop the means to include audiovisual prompts in the tests. As with most general learning platforms, testing platforms have evolved significantly in the past ten years. Some of these changes have been reflected in this journal (Roever, 2001) but, overall, the greatest challenge is not preparing the test or producing a testing platform that can test faster and more efficiently than teachers do today, but rather, first, being able to develop specific items only designed for computers and, second, studying how the results and correction of online tests can lead to learning (Chapelle & Douglas, 2006). Figure 1 shows the interface of a low-stakes testing platform where the students can see corrections. The student can clearly see different codes according to the mistake produced and, if a list of color codes is provided, will be able to recognize the type of mistake that was made. Additionally, some testing platforms include comments from the raters but it is not clear whether these comments lead to language learning.

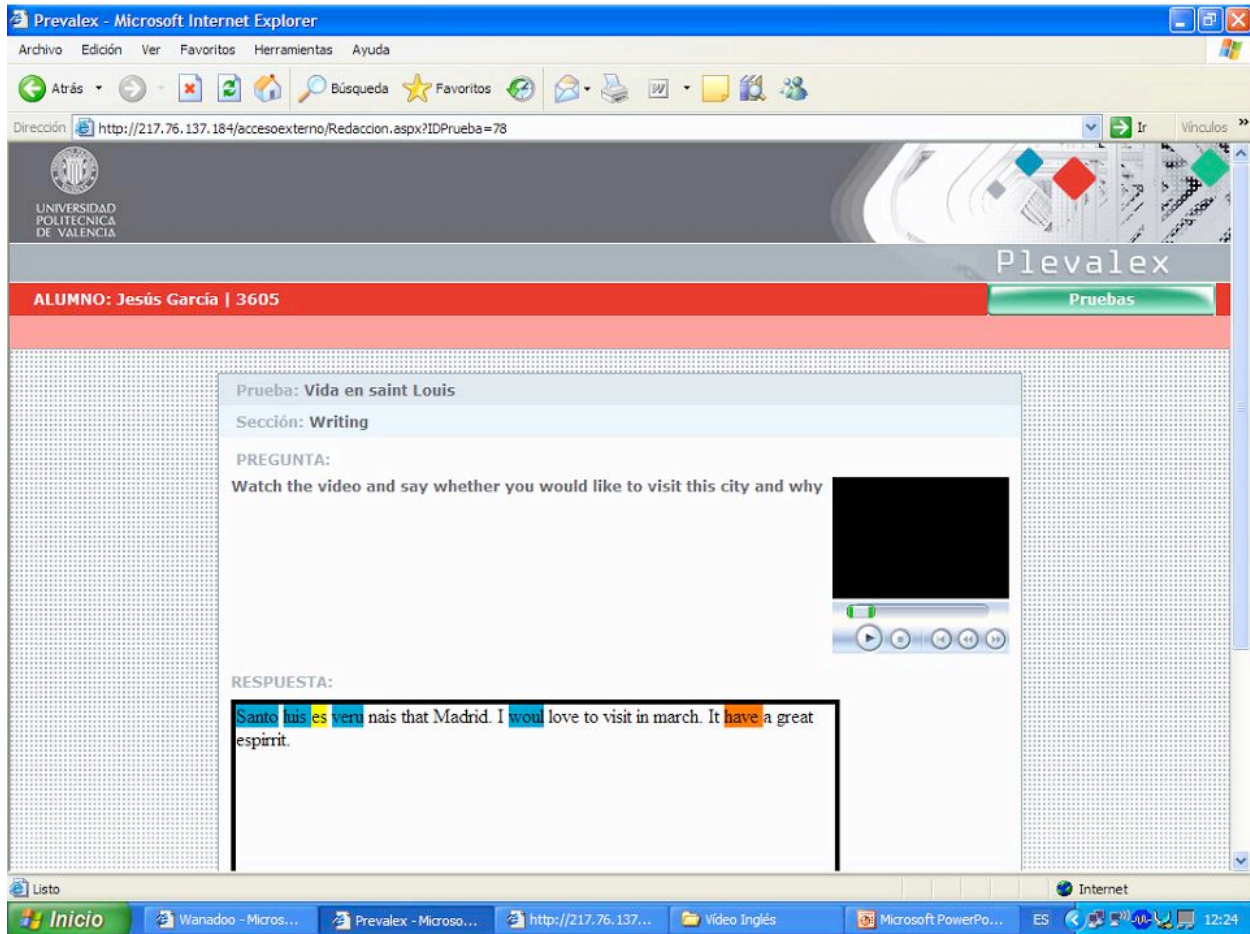


Figure 1. Interface of the PLEVALEX low-stakes testing platform (García Laborda, 2006).

The fact that it was necessary to state clearly if certain features are common or even necessary when designing computer interfaces for tests led Fulcher (2003) to provide a set of guidelines that most testing platforms follow. These guidelines relate to hardware/software configurations, navigation, page layout, toolbars and controls, text, text color, icons and graphics, multimedia, delivery systems, score retrieval, database storage, test rubrics, familiarity (and ergonomics) technology, and trialing. The guidelines are currently used by the designers of tests such as TOEFL or BULATS but very few (if any) studies have proven their validity.

TYPES OF ONLINE OR COMPUTER-BASED TESTS ITEMS

From a traditional perspective, computer-based tests are usually composed of multiple-choice questions. In the last few years, many low- and high-stakes tests have included different types of exercises that can be found in many e-learning platforms. For example, the long-standing [DIALANG project](#) (an adaptive low-stakes testing platform) includes a section where users can suggest new types of items. Today, however, while multiple-choice items have preference in high-stakes exams, low-stakes exams, like those distributed by many universities all over the world tend to include more free writing sections (Bernhardt, Rivera & Kamil, 2004). In relation to the interface organization, formats vary, although most tests follow the patterns described below:

Multiple-choice items

Traditionally set as a vertical list of questions with answers, today most exams present only one question

at a time. This allows for both a bigger font size for questions and answers as well as larger displays of any media when they occur as prompts for a question. Multiple-choice tests tend to be adaptive. Adaptive tests are based on the results (right/wrong) for each test item and have the capacity to place students in the appropriate proficiency level by using an item bank (Meunier, 1994; Brown, 1997). Generally, this type of test tends to reduce the time necessary to take the test and facilitates correction and reporting.

Writing

Interfaces for writing usually show the retrieval device and provide a box to write in the answer. The key issue in these interfaces is whether this box should be large enough to let students see most of their test or whether simply scrolling up and down is an acceptable format for writing an essay comfortably (Figure 2).



Figure 2. Interface of the writing section in PLEVALEX low-stakes testing platform.

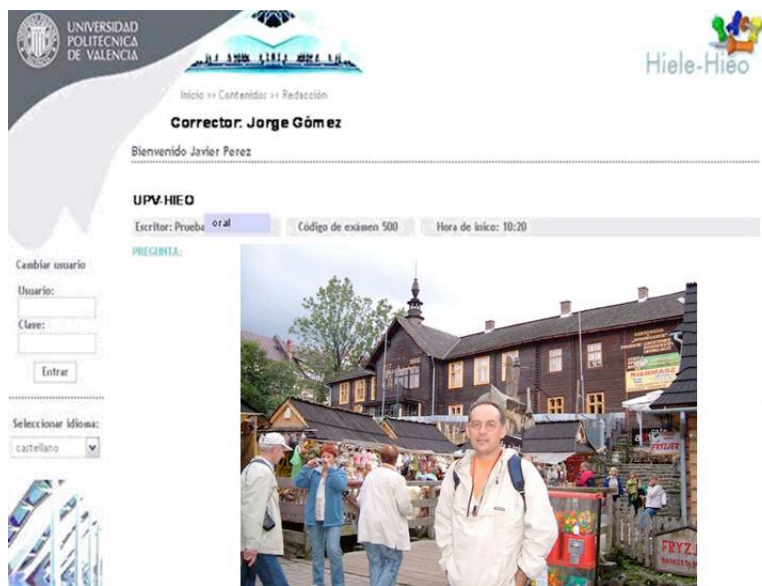


Figure 3. Interface protodesign of the speaking section in the PLEVALEX low-stakes testing platform.

Speaking

Multimedia reproduction devices usually occupy most of the screen. The idea is that the conversation should be as natural as possible. A larger picture is supposed to produce in the viewer an effect of increased identification with the task. (Figure 3)

Tests have tended to follow three types of format. Undoubtedly, these formats affect the students and the strategies used in the preparation for the test. The most significant difference is in the oral section. Some tests do not include a speaking section while others have integrated the oral section in the platform through semi-directed interviews and descriptions. Overall, there are four approaches:

1. There is no speaking section. The test includes some listening tasks, but the student is not expected to speak.
2. The platform is used for grammar, reading, and listening (and sometimes writing). There is an individually-administered speaking test administered by a human tester.
3. The testing platform is used to provide input (in the case of BEST Plus, adaptive input), but there is a test administrator present while the test takes place.
4. The speaking section is included in the computer test, and there is no test administrator specifically in charge of giving the speaking test (for example, the new iBT TOEFL).

REVIEW OF HIGH STAKES TESTS

Although recent literature has paid special attention to high-stakes tests (Stoynoff & Chapelle, 2005), very few journal publications have addressed the use of computers in high-stakes test delivery. According to some research, the difference between a paper-and-pencil test and a computer-based test is minimal (Breland, Lee, & Murak, 2005; Choi, Kim, & Boo, 2003). However, anecdotal evidence gathered from teachers at the Polytechnic University of Valencia suggests there are broad differences between students in their computer skills. Thus, there are probably some differences between the two ways of taking an exam.

This paper looks at eight high-stakes tests in order to review their features. They are classified in the following fashion: tests that only include multiple-choice questions, tests that include writing and multiple-choice questions, and tests that include speaking questions. The tests reviewed are: BULATS, IELTS, TOEFL, ACT ESL placement test, BEST Plus, Computerized English Skills Assessment (CELSA), WebCAPE (Computerized Adaptive Placement Exam), and iBT TOEFL. It is important to mention that no discrimination has been made between online tests and other computer-based tests (on CD ROM or Intranet systems) because as technology progresses, it is expected that most of these and other tests will go online.

Tests with no speaking section

ACT ESL Placement test (originally a variation of the COMPASS testing system, <http://www.act.org>) is a computer adaptive testing system for mathematics, reading and writing, and ESL. The type of questions included in the test can be found at <http://www.act.org/esl/sample.html>. For ESL, it includes sections on language use, reading, and listening that can be used singly or in combination. Additionally, Compass has developed a new writing test called ESL e-Write that will go online very soon (<http://www.act.org/compass/announce/preview.html>). This test can be used to evaluate high school and university students.

WebCAPE Computer-Adaptive Placement Exams (formerly ESL-Computerized Adaptive Placement Exam, <https://www.softstudy.com/products/CAPE.cfm>), developed by the Humanities Research Center, Brigham Young University, is a low-stakes exam usually used in commercial hiring. The test can be taken in ESL, Spanish, German, French, and Russian. The 15-20 minute test includes language use, listening, and reading and can be tailored according to the needs of the testing institution. Until recently

this test could be obtained from Brigham Young University but now is run by SoftStudy, Inc. A free version of WebCAPE is available at <https://www.softstudy.com/products/trial.cfm>.

The Combined English Language skills assessment in a Reading Context (CELSA, <http://www.assessment-testing.com/celsa.htm>) is a 75-item multiple-choice cloze test. Most institutions use it along with other spoken language tests. CELSA is used to evaluate grammar and reading at the beginner, intermediate, and advanced levels. This test can be taken in either a traditional or computer-based format. A free trial is available upon request at <http://www.assessment-testing.com/ctademo.htm>.

Computerized tests followed by a personal interview

BULATS (http://www.bulats.org/tests/computer_test.php) is a very flexible test for business Spanish, French, English, and German. Each delivery organization chooses the exact format, and results are given according to the ALTE (Association of Language Testers in Europe) Proficiency Levels (http://www.alte.org/can_do/index.php). The test uses the traditional adaptive system in which each question is presented according to the results of previous questions. A non-adaptive demonstration of this test is available at <http://www.bulats.org/demo/index.php>. The questions include listening comprehension, reading, and language use. Sometimes usability can be a little difficult because the test uses a system of arrows to proceed from one exercise to another, but, overall, it follows the format seen in the sections above. The test usually lasts about 60 minutes.

Computer Based IELTS (CB IELTS, <http://www.ielts.org/>) is a new computer version of IELTS in which the listening, reading, and writing sections are administered on a computer. However, the speaking sections are still conducted face-to-face. For the time being, this version is only available in a limited number of locations worldwide. As in some of the tests above, the listening and reading sections are rated automatically while the writing section is rated by IELTS examiners in the same manner as the paper-based version. Although IELTS has both a general and academic versions, at this moment the only available version of CB IELTS is the academic one.

The testing platform is used to provide oral input

Basic English Skills Test (BEST, <http://www.cal.org/topics/ta/best.html>) was originally designed to determine whether examinees were at the survival and pre-employment skill levels. BEST has since changed its form. The language use section of BEST Literacy is taken with pen and paper while the oral section of the BEST Plus test (<http://www.cal.org/bestplus>) is held in front of a test administrator who uses the computer as input for the test. The test takes from 5 to 20 minutes and “personal, community, and occupational domains [are] assessed using real-life communication tasks such as providing personal information, describing situations, and giving and supporting an opinion” (BESTPlus FAQ, 2007). The objective of the test is to determine whether the taker will be able to function in routine situations. In the computer adaptive version of the test, the computer is used as a prompt to provide input and set the items, and then the test administrator simply enters the item score in the computer and the computer selects the following appropriate question for the examinee's determined level.

The speaking section is included in the computer test

The Test of English as a Foreign Language (TOEFL) assesses all four skills. Listening and reading, as well as language use, follow the traditional models of computer adaptive tests. The writing module has a double correction system: automatic computer-based and human. The Educational Testing Service claims a high degree of correspondence between the two versions (Chodorow & Burstein, 2004). Input is elicited by multimedia prompts such as interviews, descriptions, and questions or other types of items. The new test will soon be available worldwide.

THE FUTURE OF ONLINE COMPUTER ASSISTED LANGUAGE TESTING

In the future, most standardized tests will go online or be based on inter- or intranet technology. They will hopefully include speaking sections. In the short term, there are some issues with technical access, test items that are not tailored to the computer platform, and the unknown effects of computerized tests on language instruction that need to be addressed.

- *Security.* Administering organizations should make sure that there will be no possibility of anyone breaking into the system and obtaining any specific version of any test.
- *Technical devices.* These tests should be adaptable to any computer environment (Windows, Mac OS, Linux, etc.) and any type of devices (PDAs, laptops).
- *Technical capability.* Even TOEFL has been reported as having technical problems in Shanghai (Guardian Unlimited, October 20, 2006). Networks need to be improved, and problems with electricity supply and Internet connection need to be solved in some instances.
- *Test features and design.* It is necessary to design new types of items for computers, especially for Internet-based tests (Davies, 2003).
- *Test interpretations and washback.* There are still very few studies of the effect of computer-based tests on how teachers change their instruction style according to the computer interfaces in standardized tests, and the results obtained in this type of exams (however, see Bailey, 1999).

There are other long-term challenges related to online testing that will also need to be examined.

- *Communities of language learners.* There are communities of test takers that could communicate test strategies and information among themselves. These communities, if culturally based, could help many test takers overcome a lack of knowledge of the Anglo-European culture that underlies most of these tests.
- *Language test preparation.* Language test preparation should also aim at improving the students' language skills; thus specific research on the impact of computer-based exams in language acquisition is still necessary.
- *Test enjoyment.* This is a very new and almost unexplored issue. Egbert (2003) has suggested that CALL be defined as a "flow move" in which the student progresses with a feeling of enjoyment that goes beyond the moment to turn into a particular state of mind. It would be interesting to research whether students experience an agreeable feeling as they progress through a test, as suggested by Egbert for CALL in general.

In final analysis, however, the benefits of online testing should overcome any of its drawbacks, as it can be faster, more efficient and less costly than traditional paper-and-pencil testing. Additionally, multimedia prompts can help make the test feel more "real." Adaptive tests can facilitate the difficult task of rapid diagnosis, and self-correcting tests can accelerate the process of correction, feedback, and reporting.

CONCLUSION

This paper has attempted to show the generic types of test items and interfaces and to present a brief review of the most widely-used ESL computerized tests. It is a first step for those who wish to take a greater interest in online language testing. The coming years will surely bring many improvements in language testing that cannot yet be foreseen.

NOTES

1. Although some researchers make a distinction, in this paper I will use *exam* and *test* interchangeably.

REFERENCES

- Bailey, K. M. (1999). *Washback in Language Testing*. ETS Monograph Series, MS-15 Retrieved May 21, 2007, from www.ets.org/Media/Research/pdf/RM-99-04.pdf.
- Bernhardt, E. B, Rivera, R. J, & Kamil, M. L. (2004). The practicality and efficiency of web-based placement testing for college-level language programs. *Foreign Language Annals*, 37(3), 356-366.
- BEST Plus Frequently Asked Questions. (2007). Retrieved May 21, 2007, from http://www.cal.org/bestplus/faqs/gen_info.html.
- Brown, J.D. (1997). Computers in language testing: Present research and some future directions, *Language Learning & Technology*. 1(1), 44-59.
- Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating *e-rater's* performance on TOEFL essays, TOEFL Research Report No. RR-73, ETS RR-04-04). Princeton, NJ: ETS. Retrieved on January 10, 2006, from <http://www.ets.org/Media/Research/pdf/RR-04-04.pdf>.
- Choi, I., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295-320.
- Chapelle, C. A., & Douglas, D. (2006). *Assessing Language Through Technology*. Cambridge: Cambridge University Press.
- Davies, A. (2003). Three heresies of language testing research. *Language Testing*, 20(4), 355-368.
- Egbert, J. (2003). A study of flow in the foreign language classroom. *Modern Language Journal*, 87(4), 499-518.
- Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, 20(4), 384-408.
- García Laborda, J. (2006). PLEVALEX: A new platform for Oral Testing in Spanish. *Eurocall Review*, 9, 4-7. Retrieved on January 10, 2006 from <http://www.eurocall-languages.org/news/newsletter/9/index.html>
- Guardian Unlimited. (October 20, 2006). ETS denies technical problems with online Toefl. Retrieved January 10, 2007, from <http://education.guardian.co.uk/tefl/story/0,,1929295,00.html>.
- Meunier, L. E. (1994). Computer adaptive language tests (CALT) offer a great potential for functional testing. Yet why don't they? *CALICO Journal*, 11(4), 23-39.
- Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, 5(2), 84-94.