ISSN 1442-438X

CALL-EJ Online, *9*(1), 15-22

From Fulcher to PLEVALEX: Issues in Interface Design, Validity and Reliability in Internet-Based Language Testing

Jesus Garcia Laborda (garcial@upvnet.upv.es)
Universidad Politecnica de Valencia, Spain

Abstract

Interface design and ergonomics, while already studied in much of educational theory, have not until recently been considered in language testing (Fulcher, 2003). In this paper, we revise the design principles of PLEVALEX, a fully operational prototype Internet-based language testing platform. Our focus here is to show PLEVALEX's interfaces and indicate their relation to Fulcher's design principles (2003). We also suggest some tentative design changes for further projects. Additionally, we will summarise how PLEVALEX benefits from the validity and reliability features of general tests delivered through computers. (9 references)

Traditionally, interfaces in language testing have not received special attention in language testing because designers of L2 platforms followed the same criteria for tests and general L2 learning platforms. In 2003, Fulcher was the first writer to formulate the basics of interface design to be taken into account in computer-based language testing. Fulcher's theory is considered today as the main theoretical framework for Computer-assisted language testing (CALT) interface design. This study evaluated the relation of the CALT PLEVALEX platform with his theories. Although the platform design has been justified by Garcia Laborda (2006a) and validated by Garcia Laborda (2006b), its design needs to be seen in the light of the current theories of which Fulcher's is foremost. The key issue behind this study is to suggest the positive features that make PLEVALEX a facilitating tool for delivering language testing. This paper is broadly divided into three sections: Introduction, Application of Fulcher's theoretical principles, and conclusion.

Introduction

PLEVALEX was developed by the GILFE group assisted by a grant provided by the Education Department of the Autonomous Region of Valencia (Spain) and the Research Incentive Program of Universidad Politecnica de Valencia between 2004 and 2006. PLEVALEX delivers tests that include oral and written tasks through the Internet. The platform enhances these tasks by using multimedia materials to support the questions. The inclusion of multimedia clues in language testing has been justified in several studies (Verhallen et al., 2006) and by the research projects currently followed by TSE in the United States. According to these studies, when conventional and CALT are compared, CALT multimedia testees seem to improve their memory recall and an increment in short term memory. Thus, when two tests with similar content and tasks are compared (except for the delivery system) the students' performance in CALT is expected to excel over traditional tests. However, the design principles of pen-and-paper tests are quite different

from those used in CALT. So far, most data showing improved benefits of CALT over traditional testing comes from general experiences in language testing such as the studies by Ngu and Rethinasamy (2006) or Munoz (2006).

Fulcher's theoretical principles on interface design

In 2003, Fulcher published a paper that established the guidelines for testing web design. After his seminal paper, there seems to have been little revision of his ideas except for Chapelle & Douglas (2006). For many researchers in testing, the key issue is whether a piece of software specially designed for language testing follows the guidelines established by Fulcher. His paper is centred around three main points:

1. Prototype design. This section is mostly devoted to describing the general specifications that both software and hardware should account for.
2. Practices in interface design. This point includes navigation, terminology, page layout, text presentation and color, toolbars and control bottoms, presentation of icons and graphics, presentation of "help" possibilities, item types, and multimedia clues inclusion.
3. Concurrent design activities. This aspect relates to technical aspects in software design such as delivery systems, score and rater's retrieval, technological and software architecture, and trialing.

Despite the importance and influence that this paper should have had, most practitioners in the field follow other guidelines more related to general educational interface design than language or even testing specific design.

Designing testing interfaces: PLEVALEX

PLEVALEX is a testing platform developed by the GILFE group at the Polytechnic University of Valencia. It has been described in previous articles (Garcia Laborda, 2006a). The platform includes three types of tasks: multiple choice (that can include grammar, listening comprehension and reading comprehension items), written (includes all the previous items plus writing short answers and essays) and finally oral tasks (semi-directed interviews, oral short and longer answers, opinion items and descriptions) (Table 1).

Table 1
Item types included in PLEVALEX (García Laborda, 2006b)

Section→	Multiple Choice	Written Composition	Speaking section
Tasks→	Choose one (correct)answer	short answers Prompted or figure aided composition	General pair directed questions Lecture and question-answer Diagram/picture description

Designing prototypes

Garcia Laborda (2006b) recently compared the iBT TOEFL platform and PLEVALEX in their goals, design, and achievement. However, it is necessary to analyse whether interface design also conforms to the basic principles of CALT design (Fulcher, 2003). To approach the interface design in PLEVALEX the researchers followed the analysis by Chapelle and Douglas (2006: 84-85) to address Fulcher's theories.

Category of the design issue	General consideration process	Application to PLEVALEX
designing prototypes	Hardware/software considerations	The platform was designed with Microsoft components although further developments will include a design with free software. Although the prototype was designed for desk PC's there is currently research in progress to adapt the software to laptops and freeware browsers.
Implications: One of the major problems in oral testing is the high cost of personal interviews or, when using computers, the licenses of authoring software. If the exams can be distributed through different means such as mobile devices they do not need to be held in computer labs or resources from different origins can be used, thus reducing costs. Further, when there is no limitation to the number of posts available for free (as a result of the number of licenses granted) the costs are reduced as well.		

Figure 1 Approach the interface design in PLEVALEX

Designing interfaces

The PLEVALEX project began its design long before Fulcher's theories on interface design were published. It is also worth mentioning that although the starting idea for PLEVALEX was planned taking into account a different set of goals and a different profile of test takers from TOEFL, which it resembles most, both share many of the interface design features (Garcia Laborda, 2006c). Fulcher's theories related to interface design were based on considerations of navigation, terminology, page layout, text inclusion, text color, toolbars and controls, icons and graphics, help facilities, item types, and multimedia inclusions. These factors are addressed step by step to make clear how the PLEVALEX platform addresses them. Figure 1 presents an interface of the actual testing platform in which most of the following ideas will be better seen.

Category of the design issue	General consideration process	Application to PLEVALFX
Good interface design (1)	Navigation	Navigation and buttons are clearly defined both by colors and also by shape and definition. The functions are written, enabling students to see them immediately. The titles and exam sections are also obvious and the icons to operate the multimedia are universally recognized.

		Students will be able to review questions within the same exam section but not to go over one that has already been done and closed.
	Page Layout & Toolbars and controls	When designing PLEVALEX there was a clear interest in limiting the ornaments and artwork that can be seen in regular educational websites. Test takers should only focus on the test and thus toolbars have been eliminated, scrolling reduced to a minimum, and upper case eliminated.
Implications: Students can focus on the exam contents because the functioning mechanisms are limited to the simplest tasks such as go on, proceed, and go to the next screen. Problems in language testing might emerge if students had too many options to operate the system. When students have extra difficulties managing a computer tool, and this difficulty might impact on the test score, the test should be invalidated. Thus, the effect of the tool should only be a context in which the language proficiency can be measured, and the results on a paper test should be equivalent to those measured on the web tool.		
Good interface design (2)	Text	As can be seen, the text is centered and limited to general and specific instructions. No additional text has been included and students are not distracted from their tasks. Font size is between 12 and 16 so it can be seen easily and the type is either Verdana or Arial to allow international recognition.
	Text color	Text colors are limited to black and white. Although at the beginning the website included certain fonts and buttons in red and green, they have been progressively eliminated to accommodate test takers who might have problems with these two colors (Daltonism).
Implications: Reading has until now been a doubly important task: firstly, because students need to read to understand the test instructions and secondly, to complete the reading tasks. Further, instructions are not usually heard by the test taker (except in the oral tasks in which a reading description is also provided). Therefore, PLEVALEX presents an optimal configuration in presenting both instructions and reading texts (see diagram 2). This may eliminate the typical constraints found in international students especially from Asia or Africa who may not be familiar with certain fonts or sizes.		

Diagram 1 PLEVALEX interface design for Spanish Speakers (Writing)

design issue	General consideration process	Application to PLEVALEX
Good interface design (3)	Icons and graphics	The interface presents very few icons. As can be seen in diagrams 1 and 2, graphics are limited to the heading, a timer, and the process buttons ("stop", "start", "next").

	Multimedia	Three possibilities were considered to present the multimedia prompts; they are presented as in diagram 1. However, in the oral section, the size can be increased up to half-screen size to increment the student's comfort and familiarity with the images. This is especially relevant in the oral section where students semi-interact with videoclips.
Implications: Although icons can be facilitating tools in textbooks, their use should be limited to avoid extra distractions. Also, the multimedia repertoire used in PLEVALEX facilitates the student's attention towards the tasks. It has also been observed in pilot studies that students engage with the images without problems.		

Figure 2 PLEVALEX interface design for Spanish Speakers (Reading)

Other considerations in interface design

Fulcher also considers other issues that this paper addresses in a limited way. These issues are related to task design, presentation, and concurrent activities. Concerning the tasks, PLEVALEX includes more than 20 types based on the combination of visuals, listening, writing, and speaking. Fulcher included in the concurrent activities the following:

Fulcher's concurrent activity	In PLEVALEX, ...
Delivery systems	Internet and Intranet
Score retrieval	Students can access a special report and score presentation screen that includes the exam rating and final grades.
Database storage	The platform uses two servers to distribute and store the data. Exams are recorded on both servers.
Test Rubrics	The tests follow the rubrics and measures that guarantee an optimal test design and distribution. Additionally, exam centers centres will have paper copies for emergency situations.
Familiarity studies	Carried out before, during and after designing the platform.
Technology studies	Performed at three levels: designer, internal validators and external validators.
Trialing	Small-scale trialing has been performed and larger-scale trialing is in progress.

How do validity and reliability improve through I-CALT?

Although some teachers tend to believe that tests measure the overall L2 proficiency of the test taker, in reality, a test only shows certain indicators of L2 proficiency. If a student is evaluated, for example, on reported speech tasks, and the student is fully aware of the type of task that will be assigned (for instance, in First Certificate, TOEFL or any other standardized test) and has developed the necessary strategies to optimize the results, the grades could be better than those obtained by another student who has not done any specific training in those reported speech tasks. Thus, content or language forms in an exam are just as vital as how they are presented. A long-standing question in language testing is how to infer what the exam administrators (either a class teacher or an agency such as the Cambridge Board) want to measure and if what is measured shows what is intended to be measured (for instance an advanced proficiency level in the CAE test). Sometimes, achieving a match between these two is so difficult that Alderson, Clapham & Wall (2001) have suggested the near impossibility of writing good tests.

PLEVALEX, as well as other I-CALT tools, includes prompts that facilitate the identification of the task with real language knowledge and use. Figure 1 asks the student to see and understand the video clip and then write a response. A traditional writing task becomes a complex language exercise in which listening and writing are measured but also measured is the student's capacity to process visual information, to understand the outside world, and to conceptualize all this information in a writing task. In this way, some problems are created by the selection of adequate reading texts or writing topics. Visual prompts, as opposed to listening texts, usually allow a certain degree of variability in the test taker's understanding given that, although the student in visual prompts may not fully understand the audio message, the supporting images can still provide some ideas and support that facilitate the student's response. Additionally, the supporting images facilitate the student's transition between easy and difficult listening and writing sections through the inclusion of vocabulary or expressions suggested by the video clips not included in the listening or reading text. Thus, tests which have audiovisual prompts delivery, such as in PLEVALEX, are "more real" and tend to measure the complexity of the language better.

If a computer tool may be valid and reliable for multiple-choice and writing tasks, the new challenge undertaken by PLEVALEX was to include semi-directed interviews to assess oral proficiency. In traditional face-to-face tester? testee contexts it may be difficult to establish procedures to evaluate the student's oral skills. The difficulties lie in the type of language expected, topic appropriacy, and the test method. PLEVALEX has been designed for three types of tasks: descriptions, semi-directed interviews, and responses to opinions or information questions (not in a conversation). Although some difficulties may arise in those students who may not be familiar with the test format, trials in the Polytechnic University of Valencia have proved that students can deal properly with three types of tasks.

PLEVALEX database facilitates the reliability of the grading system because raters have the opportunity to revise both written and oral tasks. Video recordings also permit consideration of non-linguistic features better than through a simple audio recording. PLEVALEX permits both audio and, shortly, video recording of testees' responses. Databases also permit a refining process though a continuous reevaluation of proficiency standards due to a continuous revision of responses. The reliability of the scores can also improve because testers do not need to grade the test-taker's performance while the test

is being administered, as opposed to face-to-face tests where decisions are taken immediately. Despite the continuous training, some tests have proved that differences among raters exist and part of these differences may diminish through the use of databases and differed correction. Therefore, raters have the opportunity to revise their work (and mark it twice or revise their correction) and adapt their criteria to the common principles of the test, thus reducing the differences among raters and reducing inter-rater discrepancies that may affect reliability. Also, for the speaking section, PLEVALEX offers the possibility to use a test giver and a rater so the "optimal" corrector does not need to be a good input provider. This means that the recording can be "acted" by an articulate speaker with neat and optimal speech who will not necessarily have to correct the test (Alderson et al., 2001).

Conclusions

The interface design meets the most important requirements suggested by Fulcher in his seminal paper. Although further applied studies are required by the PLEVALEX testing platform, the adequacy of PLEVALEX's current design and presentation in terms of Fulcher's framework, its validity and reliability seem evident. The designers will have to consider the evolution of the current TOEFL and IELTS platforms as well as their research. Overall, the research and design teams, comprising more than 15 university professors and researchers, suggest PLEVALEX could be a tentative step towards facilitating solutions to many of the current problems in online and general low stakes language testing.

References

Alderson, J. Clapham, C. & Wall, D. (2001). *Language test construction and evaluation* (pp. 62-64). New York: Cambridge University Press.

Chapelle, C.A. & Douglas, D. (2006). *Assessing language through computer technology*. New York: Cambridge Univ. Press.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language testing*, 20 (4): 384-408.

Garcia Laborda, J. (2006a). PLEVALEX: A new platform for oral testing in Spanish. *Eurocall review*, 9: 4-7. http://www.eurocall-languages.org/news/newsletter/9/index.html.

Garcia Laborda, J. (2006b). "Analizando criticamente la selectividad de ingles? Todos los estudiantes espanoles tienen las mismas posibilidades?" *Tesol Spain*, 30.3: 9-12.

Garcia Laborda, J. (2006c). "?Que pueden aportar las nuevas tecnologias al examen de selectividad de ingles? Un analisis de fortalezas y oportunidades". *Revista de CC educacion*, 206: 151-166.

Munoz, B., Magliano, J. P., Sheridan, R., & McNamara, D. S. (2006). "Typing versus thinking aloud when reading: Implications for computer-based assessment and training tools". *Behavior research methods,* 38(2), 211-217.

Ngu, B. H., & Rethinasamy, S. (2006). "Evaluating a CALL software on the learning of English prepositions". *Computers and education*, 47(1), 41-55.

Verhallen, M. J. A. J., Bus, A. G., & de Jong, M. T. (2006). "The promise of multimedia stories for kindergarten children at risk". *Journal of educational psychology*, 98(2), 410-419.