Diseño y validación de la plataforma PLEVALEX como respuesta a los retos de diseño de exámenes de lenguas para fines específicos

Jesús García Laborda y Teresa Magal Royo
Universidad Politécnica de Valencia
jgarcia@upvnet.es y tmagal@degi.upv.es

Resumen

Este artículo describe las necesidades de diseño y validación que llevaron al desarrollo de la Plataforma de Evaluación Valenciana de Lenguas Extranjeras (PLEVALEX) para exámenes orales y escritos a través de Internet. Aunque la plataforma es multilingüe, en este trabajo nos referiremos a su uso para exámenes de inglés con fines específicos. La plataforma fue desarrollada por el Grupo de Investigación en Lenguas para Fines Específicos (GILFE) de la Universidad Politécnica de Valencia. La validación se hizo en cooperación con el Centro de Investigación en Tecnologías Gráficas (CITG) de la misma universidad. La prueba de ergonomía se realizó a tres niveles: diseñadores, expertos y usuarios. El seguimiento realizado con estudiantes de comunicación audiovisual y profesores de la Universidad Politécnica de Valencia indicó una reacción positiva hacia la plataforma de exámenes en red.

Palabras clave: exámenes, lenguas de especialidad, evaluación, plataformas de exámenes, Internet.

Abstract

Design and validation of the PLEVALEX platform as a response to the challenges in the design of exams within Languages for Specific Purposes

This paper outlines the design and validation needs that led to the development of the PLEVALEX Internet-based testing platform for oral and written exams. Although this is a multilingual platform, in this paper we will focus in its use for tests for English for Specific Purposes. The platform (whose acronym stands for *Plataforma de Evaluación Valenciana de Lenguas Extranjeras*) was developed by the

Research Group in Languages for Specific Purposes (GILFE) at the Polytechnic University of Valencia, Spain. The validation was done in cooperation with the research Center in Graphic Technology from the same university. The Ergonomic test was conducted at three levels: designers, experts and users. Students and faculty from the Polytechnic University of Valencia follow up showed a positive reaction to the web-based testing platform.

Keywords: exams, languages for specific purposes, assessment, testing platforms, Internet.

1. Introducción

Vivimos una Europa cambiante en la que los estudiantes tienden a realizar una parte significativa de sus estudios en el extranjero. Los estudiantes universitarios son conscientes de los efectos positivos que tiene realizar una estancia prolongada (de seis o más meses) para conocer nuevas culturas, mejorar su segunda o tercera lengua y, posiblemente, establecer amistades o contactos que enriquezcan su futuro profesional. Asimismo, aunque menos percibido por la opinión popular, también aprenden a enfocar sus conocimientos o a adquirir otros nuevos, como nuevas formas de investigación o cierta normativa profesional de su competencia, a través del crisol de culturas y lenguas diferentes. Por supuesto, esto conlleva el conocimiento y aprendizaje de una lengua extranjera. Sin embargo, puesto que su experiencia va a ser una combinación de aprendizaje general y de la especialidad en la que se mueven académicamente, es necesario establecer mecanismos de control lingüístico tanto de diagnóstico como de evolución y seguimiento para poder optimizar su aprendizaje.

Desde siempre se han realizado exámenes de diagnóstico de la competencia en lengua extranjera de los estudiantes internacionales. Este diagnóstico es una operación especialmente costosa en recursos humanos y económicos ya que la mayoría de las universidades se ven obligadas a realizar una inversión en personal de vigilancia, fotocopias, correctores, vigilantes de exámenes, administradores de exámenes, etc. Sin embargo, la cuestión es que las universidades, conscientes de estos gastos, limitan el tiempo y formato de los exámenes para hacerlos lo más sencillos posible. Esto implica, en no pocos casos, que dichos exámenes se simplifiquen al máximo tanto en tareas como en longitud y ello hace que las universidades utilicen pruebas con tareas de selección múltiple, mucho más rápidas y fáciles de corregir. Aunque autores contemporáneos han defendido el uso de dichas pruebas como indicadores

válidos (Herrera Soler, 2005), y esta posición tiene cierta tradición histórica al amparo de la Teoría Unitaria (Oller, 1979 y 1983) y cuya limitación es evidente cuando se establecen como el único tipo de prueba. Es más, resultan casi impensables separadas de otras como redacción o comprensión oral. Sin embargo, en muchos casos todavía hoy son el único tipo de ítem incluido en los exámenes de diagnóstico y aún de asignaturas de lenguas abiertas a los estudiantes generales de cada universidad. Por consiguiente, la cuestión es, primero, si no se deben incluir otros tipos de tareas y, segundo, cómo hacerlo de una manera eficaz para no sobrecargar más el trabajo de los profesores, permitiendo a la vez medir el conocimiento de lenguas de especialidad. Así pues, es necesario encontrar un método por el que los estudiantes puedan ser diagnosticados, baremados y evaluados de una manera efectiva, aprovechando los escasos recursos que poseen las universidades (García Laborda y Bejarano, 2005).

2. Desafíos actuales en el campo de la evaluación para lenguas de especialidad

Evidentemente, existe la necesidad de realizar tareas de medición de competencias en lengua inglesa para los fines mencionados en la introducción, pero además existen otras muchas cuestiones actualmente a debate que afectan al diseño de nuevos sistemas de evaluación. En una publicación reciente (García Laborda, 2006b) se planteaban algunas de estas cuestiones como el papel del profesor, los tipos de pruebas u otros asuntos que se han reflejado en los trabajos de varios grupos de investigación en España y en el extranjero. Tal y como indicaba García Laborda (2006b), algunos de esos factores afectan directamente al diseño de los exámenes mismos, tales como el tipo de ítems (tareas a realizar en un examen) o la presencia o ausencia de componentes audiovisuales para acompañar a un ítem (por ejemplo, el efecto de incluir o no un vídeo en una prueba de redacción o de comprensión oral). Otros factores inciden posteriormente en la enseñanza (es el denominado *washback effect*, o también *backwash effect*). Especialmente significativo cuando se trata de lenguas de especialización es el conocimiento previo de la materia con la que se relaciona la lengua (por ejemplo, economía en inglés para economía). Es más, en algunos casos habría que tener cuidado con que ese conocimiento previo no sustituya el contenido lingüístico que se evalúa (como sucedería en el caso de que una respuesta pudiese ser conocida por el evaluado a través de sus estudios y no

de la información (o *input*) provista en los ítems del examen). Con el fin de simplificar las actuaciones, seguiremos un proceso lineal (véase el gráfico 1) a la hora de realizar una revisión del estado actual de la cuestión.

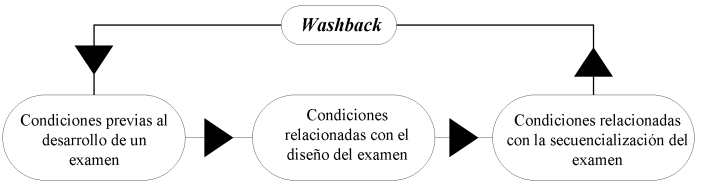

Gráfico 1. Fases de estudio de la problemática actual en el diseño de exámenes.

Hoy por hoy, al contrario que en los exámenes de criterio como el del *Test of English as a Foreign Language* (TOEFL) o el *International English Language Testing System* (IELTS), la mayor parte de los exámenes los realizan los mismos profesores de la institución. Sin embargo, en raras ocasiones estos docentes son profesionales de la evaluación y las cuestiones de diseño y seguimiento generalmente les son ajenas. Por su experiencia, filosofía de la enseñanza, evolución del aprendizaje o intereses y conocimiento de las necesidades de sus alumnos, los profesores son, potencialmente, unos evaluadores precisos. Probablemente sea porque incorporan unos conocimientos no mensurables que valoran inconscientemente, como evaluación de la personalidad y potencialidad del alumno, junto con el resultado mismo del examen que le lleva a asignar la nota final del estudiante. Sin embargo, aunque no dudamos de su buena voluntad y de la importancia de una experiencia adquirida con los años, es natural que los exámenes de diagnóstico, que a la postre posiblemente afectarán de alguna manera a los estudios y al crédito concedido a los estudios en el extranjero (*high stakes exams*), sean realizados por auténticos profesionales de la evaluación por escasos que éstos sean en nuestro país. Esta objetividad es la que precisamente debe marcar la diferencia entre el uso de exámenes en el aula y aquellos que son externos a la misma, o el paso de la subjetividad relativa de la evaluación realizada en el aula a los conocimientos mostrados a través de una prueba de evaluación objetiva de exámenes estandarizados (como por ejemplo IELTS, TOEFL, *First Certificate in English* (FCE), etc.). Sin embargo, a pesar del esfuerzo que se pueda realizar para objetivar los exámenes, no se debe olvidar que las pruebas de evaluación en lenguas de especialidad miden los conocimientos lingüísticos en un contexto profesional. Eso puede llevar a una mediación de los exámenes ya que es posible que los estudiantes que tengan buenos

conocimientos de los contenidos de las materias que se recogen en la prueba (por ejemplo, en una lectura sobre una materia que hayan podido estudiar en clase de la materia específica y en una primera lengua) consigan unas puntuaciones mejores que alumnos con mayor competencia en lengua inglesa, pero que no han estudiado dichos contenidos (Chen y Graves, 1995). En efecto, como muestra Krekeler (2006), los conocimientos previos sobre un tema afectan muy claramente al resultado de los mismos. Si esto es evidente en un tipo de tarea en la que se provee la mayor parte de la información objetiva como es la lectura, no lo es en otras en las que no se suministra, como por ejemplo la redacción (Alderson y Urquhart, 1985; Alvermann y Hynd, 1989; Chen y Graves, 1995; Fulcher, 1999; Salmani-Nodoushan, 2003) o en la comprensión oral (Jensen y Hansen, 1995). Por tanto, los exámenes de lenguas de especialidad tienen una dificultad añadida ya que deben centrarse en el área de estudio, y además deben tener suficiente información básica para garantizar la igualdad de condiciones de todos los alumnos. Aunque no hay estudios anteriores al respecto, creemos que parte de esta labor de ayuda puede hacerse mediante la presentación de ayudas audiovisuales (Anglin et al., 2002) incluidas en los exámenes. Además, lo mismo que en el caso del uso de materiales reales en la clase de lengua inglesa, es necesario que los materiales usados en los exámenes sean también extraídos de contextos profesionales auténticos (Lewkowicz, 2000; Spence-Brown, 2001).

La inclusión de materiales reales en el esquema de la prueba incide directamente en el diseño de unos exámenes que, además, deben tratar de mostrar una filosofía tradicional en la enseñanza de lenguas de especialidad, como priorizar las necesidades de lengua y contenido de los alumnos (Folse, 2006) y, así mismo, esa prioridad debe reflejarse también en el diseño de exámenes de lenguas de especialidad según propuso anteriormente Jackson (2005). En lo referente al diseño de las pruebas y tareas, muchos estudios han planteado la necesidad de romper con la tradición de exámenes basados solamente en pruebas de selección múltiple y respuesta corta. Aunque es cierto que algunos incluyen algún tipo de redacción, el hecho es que desde mediados de los 80 se ha notado la necesidad de incluir todas las destrezas lingüísticas. En nuestra opinión, el punto débil de muchos exámenes se encuentra en la carencia de pruebas orales debido a su alto coste. No valoraremos el planteamiento teórico que incide en el uso de tanto examen "objetivo" desde la década de los 80, la famosa y autorechazada Teoría Unitaria de Oller (1983). Bien al contrario, sin rechazar este tipo de tareas (apoyadas por investigaciones como las de Herrera Soler, 2005), bastantes

estudios han señalado la necesidad de incluir tareas orales en las pruebas de diagnóstico (Fulcher y Marquez Reiter, 2003). Estas tareas orales podrían ser de tres tipos fundamentales:[1]

1. La entrevista.
2. Respuestas a preguntas sociales o formales-académicas.
3. La descripción (más o menos larga) (Douglas y Smith, 1997; Folse, 2006).

En cuanto a los efectos que tienen los exámenes de lenguas extranjeras en los estudiantes, cabría destacar los llamados efectos sociales como el significado que adquieren ciertos exámenes en la vida de los estudiantes (pensemos, por ejemplo, en TOEFL en el que la obtención de una puntuación mínima significa directamente el acceso o no a los estudios superiores en los Estados Unidos). Una segunda cuestión sería cómo los resultados de los exámenes afectan a los cambios en la metodología de la enseñanza a nivel universitario en España (García Laborda, 2004) o a nivel internacional.

Aunque no es un aspecto prioritario de este trabajo debatir sobre la aplicación de ciertos métodos y tipos de tareas, sí conviene mencionar el hecho de que los cambios en el diseño de exámenes como TOEFL o IELTS suelen conllevar también cambios en la forma en las estrategias de aprendizaje y en las estrategias utilizadas para realizarlos (Alderson y Hamp-Lyons, 1996; Jamieson et al., 2000; Kobrin y Young, 2003; Cumming et al., 2005), que implican otros efectos éticos relacionados con la valoración del propio examen (Hamp-Lyons, 1998). Estos cambios son especialmente significativos cuando afectan a la forma del sistema de distribución del examen (Wise y Plake, 1989; Kobrin y Young, 2003). Un hecho que ya se ha observado y analizado en exámenes con gran impacto social (*large stakes exams*) como TOEFL, IELTS o *Business Language Testing Service* (BULATS), entre otros, conforme han ido adaptando su formato para ir poniendo en funcionamiento los exámenes asistidos por ordenador.

3. Recursos efectivos de evaluación

En un reciente trabajo de Chapelle y Douglas (2006) se señala que el uso de ordenadores no ha demostrado su efecto directo en la adquisición de

lenguas. Esta afirmación resulta sorprendente habida cuenta de que parece haber una necesidad de afirmar las bondades de la informática en la pedagogía contemporánea. Sin embargo, para dichos autores, los ordenadores y su uso en pruebas de evaluación aportan una mayor eficacia en el manejo, recogida y tratamiento de datos a través de Internet y en la transmisión y recopilación de los mismos. Evidentemente, no es una cuestión menor el hecho de que, hoy por hoy, procesos de transmisión de datos permitan de manera casi inmediata la maniobrabilidad y velocidad suficiente como para poder descargar vídeos o grabar sonidos y depositarlos en bases de datos que, a su vez, permitan ser vistos, oídos, elaborados e incorporados a exámenes en red.

Para tratar de responder a algunos de los retos expresados en la sección anterior, se pide que los métodos más novedosos de evaluación asistida por ordenador tengan en cuenta dos aspectos: capacidad de economizar los recursos, y una compilación de repertorios audiovisuales que sirva para evaluar con mayor precisión a los alumnos en sus campos de especialidad. Por ejemplo, supongamos que deseamos que un alumno de una especialidad de ingeniería nos describa o realice una pregunta relacionada con una parte específica de un mecanismo afín a su campo. Sería ideal poder pedirle que describiera detalladamente un vídeo o esquema de un objeto o mecanismo en funcionamiento o desde varias perspectivas, algo que raramente se puede hacer con un simple diagrama en un examen de papel y bolígrafo (Kobrin y Young, 2003). De esta manera, se añade un matiz de realismo que con exámenes tradicionales es muy difícil de alcanzar, que se ajusta a las realidades comunicativas que los profesionales encuentran en su entorno laboral (Connelly, 1997; Jacoby y McNamara, 1999) o preparación profesional (Yoshida, 1998). Por tanto, es necesario desarrollar herramientas informáticas eficientes que incluyan secciones orales y escritas y registros formales e informales genéricos, ya que el centrarse, por ejemplo, en registros de su misma especialidad daría unos resultados limitados porque el conocimiento del contenido de un examen no garantiza un conocimiento de la lengua inglesa, aunque sí un buen resultado en exámenes específicos (Salmani-Nadoushan, 2003; Cargill, 2004). Por ejemplo, un alumno podría comprender el movimiento de un motor y explicarlo mejor que el que el alumno que no lo conoce aunque el segundo tuviera una mayor competencia en la lengua inglesa Por tanto, en la medida de lo posible se debe tratar de eliminar la mediación introducida por el conocimiento de una determinada materia.

García Laborda y Bejarano (2005) analizaron las necesidades lingüísticas de los estudiantes de lengua inglesa con fines específicos a su llegada a grandes universidades como la Universidad Politécnica de Valencia o Valdosta State University (Georgia, Estados Unidos). El mismo trabajo apuntaba la necesidad de desarrollar nuevas plataformas para antes de su llegada a esas universidades y evaluar los conocimientos de la segunda lengua en los contextos de especialidad. A lo largo de unos 30 años se han ido introduciendo plataformas de exámenes de lenguas de especialidad pero muchas han quedado ya obsoletas y otras eran muy incompletas tal y como cita Roever (2001). De 2003 a 2006 el Grupo GILFE de la Universidad Politécnica de Valencia ha desarrollado una plataforma que está en su tercera fase de pruebas y, por tanto, en un momento crucial en su fase de validación. La validación de plataformas de examen, como los mismos exámenes, es un proceso complejo que conlleva gran cantidad de pruebas y medidas y, además, exige la participación de equipos interdisciplinares. Asimismo, cuando hablamos de plataformas informáticas, esta dificultad se incrementa enormemente. PLEVALEX, la plataforma informática desarrollada a partir del proyecto de la Herramienta Informática para Evaluación de Lenguas Extranjeras (HIELE) y del proyecto de la Herramienta Informática de Evaluación Oral (HIEO), ha requerido dos grupos de investigadores para su validación: el lingüístico-pedagógico y el de diseño informático. En el terreno lingüístico, la validación llevada a cabo se basa en el cumplimiento de la funcionalidad y usabilidad del entorno creado es decir, si PLEVALEX es válida para evaluar los conocimientos en lengua inglesa para fines específicos de sus usuarios de acuerdo a los parámetros propuestos por Fulcher (2003). La plataforma fue diseñada para baremar de 150 a 200 alumnos simultáneamente en sus destrezas lingüísticas orales y escritas y en su conocimiento de la gramática.

3.1. Descripción básica de la plataforma PLEVALEX

La plataforma informática PLEVALEX consta de tres módulos fundamentales, descritos con más detalle por García Laborda (2004 y 2006a). En toda la plataforma las tareas pueden apoyarse en un repertorio de imágenes y grabaciones de audio que facilita, como se ha indicado anteriormente, un mejor resultado en el examen. Un módulo se dedica a la evaluación de las destrezas gramaticales a través de preguntas de selección múltiple. El segundo módulo es de escritura y el tercero es un módulo para exámenes orales semipresenciales y semidirigidos. El primer módulo corrige las tareas de forma automática, mientras que los módulos dos y tres se

corrigen tras ser enviados los resultados a los correctores que realizan el trabajo en su lugar ordinario de trabajo. Los resultados numéricos son enviados a una base de datos que después los clasifica y ordena. Las pruebas, de esta manera, quedan agrupadas y guardadas, aumentando su control y su disponibilidad inmediata.

4. Validación de plataformas lingüísticas *online*

García Laborda y Enríquez Carrasco (2005) y García Laborda (2006a) han descrito ampliamente PLEVALEX como la primera plataforma de exámenes orales y escritos integrados que está llamada a ser una herramienta fundamental en los exámenes de alto impacto (*high stakes)* en España. De hecho, para García Laborda (2006b), un uso significativo podría ser una hipotética Prueba de Acceso a la Universidad (PAU) asistida por ordenador, ya que por primera vez permitiría una evaluación oral y escrita global y completa.

La validación de plataformas *web* educativas como PLEVALEX suele efectuarse a tres niveles que van de un proceso de constructor a usuario. De cualquier manera, esta validación se hace considerando dos factores fundamentales: su utilidad lingüística y su "usabilidad" (término recién incorporado a la jerga de la informática). Para ver mejor este proceso, obsérvese el gráfico 2.

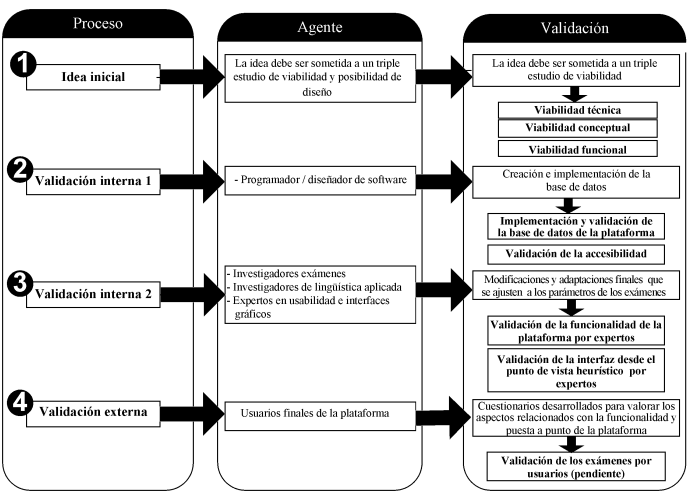

Gráfico 2. Proceso de validación de la plataforma de evaluación PLEVALEX.

Curiosamente, y puesto que aunque los orígenes de esta plataforma son simultáneos pero totalmente diferentes de los que originan el examen TOEFL, su parecido formal es sustantivo por lo que algunas publicaciones sobre el nuevo *Internet based TOEFL* (iBT TOEFL) (Zareva, 2005) son aplicables, en gran medida, a PLEVALEX.

4.1. Idea inicial: validación de los tipos de ejercicios

La plataforma plantea niveles de multitarea libre. Así pues, el alumno es capaz de poder elegir qué tareas realiza y cuándo: redacción, examen oral o selección múltiple (que incluye comprensión oral similar a la recientemente incorporada por la plataforma de diagnóstico de competencia en lenguas extranjeras *Dialang*[2], ejercicios tradicionales de selección múltiple y otras tareas similares). Esto se hace mediante un interfaz de selección de tarea que da independencia al estudiante para elegir lo que desea. Todos los ejercicios están apoyados (cuando así lo desea el examinador) por elementos audiovisuales.

4.2. Validación técnica de la plataforma

La validación técnica de la plataforma se ha centrado en una evaluación continuada del programa informático a medida que se ha desarrollado la aplicación final, siguiendo protocolos de evaluación tanto a nivel interno, sobre la base de datos en sí misma, como a nivel externo, a través de las pruebas de uso realizados a los usuarios, ya sean alumnos o profesores. Anteriormente se habían planteado validaciones por parte de expertos en los diferentes estadios de creación de la aplicación que permitieron resolver posibles errores en su funcionalidad. En la actualidad la plataforma se halla en su validación final a través de las pruebas de uso realizadas a los usuarios, ya sean alumnos o profesores.

La necesidad de una valoración "única de usabilidad" de cualquier herramienta telemática interactiva es fundamental para poder validarla frente a sus futuros usuarios, pero no es suficiente. Existen numerosos métodos de evaluación centrados en el uso de evaluadores con perfiles diversos (alumno, corrector, administrador, etc.) de los cuales se obtienen diferentes puntos de vista de la plataforma bien de forma global (manejo del entorno de la herramienta o de su interactividad) o de forma específica (seguimiento y manejo de tareas específicas, validación de tareas, etc.). De hecho, para la plataforma PLEVALEX se utilizaron tres métodos de evaluación de

usabilidad en el sentido que se explica a continuación y tal y como se ilustra en el gráfico 3:

1. La evaluación *Cognitive Walkthroughs*. Método desarrollado por Lewis (Lewis et al., 1990), en el que se simulan los problemas de los usuarios de forma escalonada y pormenorizada, sobre todo, analizándolos paso a paso, tarea a tarea, desde un punto de vista cognitivo. Para ello, se utilizan perfiles de usuarios expertos. Su aplicación se realiza en los primeros estadios de formalización del programa, herramienta o entorno telemático cuando aún es simplemente un prototipo.
2. La evaluación heurística. Es un método planteado durante los años 90 por Nielsen y Molich (1990) y Nielsen y Mack (1994), en el que un experto aplica unos principios o criterios de usabilidad sobre el programa, herramienta o entorno telemático ya desarrollado. En la actualidad existen numerosos criterios categorizados que han llevado a plantearse hasta 294 posibles problemas (Nielsen, 1994). Sin embargo, al evaluar la herramienta, y en base a una necesidad más rápida y operativa de los criterios heurísticos más generalizados, nos centramos solamente en 10 de los niveles de Nielsen (1994) y otros dos propuestos por nosotros mismos, los cuales fueron específicamente categorizados para la plataforma PLEVALEX.
3. Prueba de usuarios convencionales. Se hicieron siguiendo diversas metodologías de análisis de plataformas de entornos telemáticos creados o adaptados intencionadamente según la necesidad de valoración (funcionalidad, ergonomía visual, manejo de contenidos, etc.) (Jeffries et al., 1991). Los perfiles de usuario son bastante amplios y, según sus circunstancias, se valoran o no los conocimientos previos sobre el soporte y/o la plataforma.

La plataforma PLEVALEX utilizó los tres métodos expuestos para validar cada una de las fases determinantes de la aplicación centrándose en el cumplimiento de tres requisitos fundamentales:

- La resolución de problemas relacionados con la usabilidad de la aplicación PLEVALEX a nivel técnico, para resolver los problemas relacionados con el manejo, comprensión del proceso y método de realización.

- La utilización de criterios heurísticos para guiar los conceptos a desarrollar en los informes elaborados por los diferentes expertos de las disciplinas involucradas (informática y filología fundamentalmente). Este punto es importante ya que se les pedía a los expertos que realizaran una valoración en base a una serie de criterios heurísticos comunes que fueron analizados y comparados de forma global.
- La validación ergonómica y de funcionalidad de la aplicación PLEVALEX sobre usuarios, mediante la aplicación de la prueba a estudiantes con un perfil cognitivo y tecnológico específico afín al de los estudiantes que sean usuarios de la plataforma en el futuro y que, al mismo tiempo, sean capaces de comprender el medio informático evaluado, valorar su uso y realizar una estimación de medios telemáticos interactivos incluidos en el mismo.

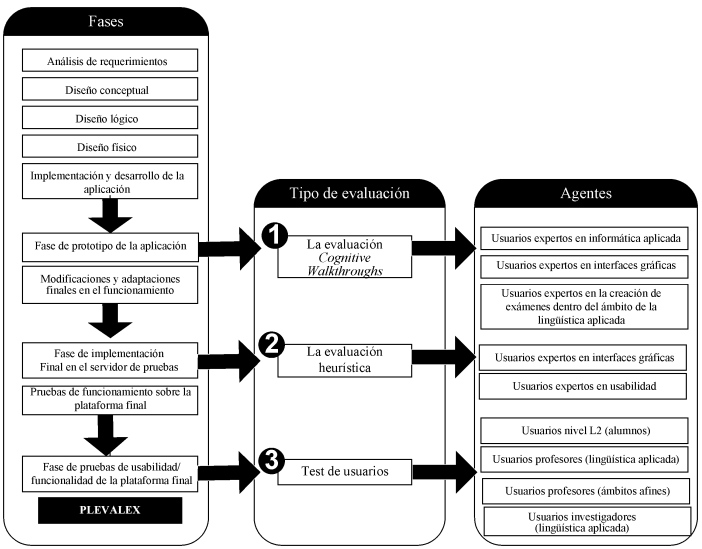

Gráfico 3. Fases y tipos de validaciones de la plataforma PLEVALEX.

4.3. Planteamiento del método de evaluación heurístico

La evaluación heurística se utiliza generalmente para evaluar herramientas o plataformas *web* en fase final antes de su puesta en marcha de forma real. Este tipo de evaluación se realiza antes que las pruebas de usuarios, porque

es capaz de detectar problemas graves de funcionamiento de la plataforma que el usuario convencional no detectaría. En comparación con otras técnicas de evaluación de interpretación de las acciones de los usuarios como la evaluación *Cognitive Walkthrough*, en la evaluación heurística no es necesaria una interpretación externa, ya que la información se halla contenida en los informes realizados por los evaluadores expertos. El procedimiento general del método obliga a los evaluadores seleccionados a inspeccionar la plataforma individualmente y a emitir informes por escrito. Las sesiones de evaluación duran aproximadamente entre una y dos horas por página. Los evaluadores utilizan una lista de criterios siguiendo las pautas heurísticas definidas previamente, que pueden considerarse de carácter estándar o que se establecen en función de las características de la herramienta.

Principio heurístico	
1	Visibilidad del estado del sistema.
2	Concordancia entre el lenguaje del usuario y el mundo real.
3	Control del usuario.
4.1	Consistencia y aplicación de estándares: nivel técnico y formal.
4.2	Consistencia y aplicación de estándares: nivel de contenidos educativos.
5.	Prevención de errores.
6.1	Reconocimiento más que recuerdo: adaptación al medio y al proceso.
6.2	Reconocimiento más que recuerdo: realización de las pruebas y elaboración de contenidos.
7.	Flexibilidad y eficacia de uso.
8.	Diseño atemporal y minimalista.
9.	Ayudas al usuario, diagnosis y recuperación frente a posibles errores.
10.	Documentación informativa adicional y acceso a la ayuda del programa.

Tabla 1. Selección de criterios heurísticos aplicados a la plataforma PLEVALEX.

Los criterios heurísticos elegidos para valorar la plataforma PLEVALEX fueron los 10 criterios básicos planteados por Nielsen (Nielsen y Molich, 1990; Nielsen, 1994; Nielsen y Mack, 1994), junto a una selección de categorías secundarias específicas que se consideraron importantes (véase la tabla 1). Como ya hemos comentado anteriormente, esta selección sirvió de punto de partida para cada uno de los evaluadores participantes.

4.4. Planteamiento del método de evaluación basado en pruebas heurísticas

En un método tradicional de uso, la mayoría de las pruebas de usuario, disponen de información controlada sobre el proceso a realizar, permitiendo en muchos casos el comportamiento espontáneo sobre la plataforma y en el descubrimiento del manejo del entorno de forma intuitiva. Una evaluación de usuario debe utilizarse siempre después de otro tipo de evaluación más controlada por expertos, como es el caso de la evaluación heurística

desarrollada para la aplicación PLEVALEX. En caso contrario, una evaluación heurística a posteriori de una prueba de usuario sólo serviría para detectar problemas superficiales que podrían haberse evitado invirtiendo el proceso.

Los perfiles de usuario para validar la plataforma PLEVALEX se centraron en la búsqueda de perfiles de alumno como agente implicado en el desarrollo de las pruebas creadas, y perfiles de profesores relacionados directamente o indirectamente con el desarrollo y gestión de pruebas de manera interna. En los dos casos, se planteaba la necesidad de evaluar las dos herramientas de la plataforma, denominadas *Backoffice* (parte del profesor, corrector o administrador) y *Frontoffice* (parte del alumno). La elección de los estudiantes dependió de un nivel mínimo de comprensión de inglés (ya que disponían de dichos conocimientos previos procedentes de la educación secundaria española y del aprendizaje adquirido de forma personal). Las variables aplicadas en la creación de las preguntas de la prueba se centraron en varios aspectos:

- Tiempo de realización de la tarea y/o proceso.
- Memoria de reconocimiento del entorno que ayude a establecer un criterio intuitivo a la hora de realizar el examen.
- Grado de satisfacción en el uso de la herramienta debido a su entorno.
- Grado de satisfacción en el uso de la herramienta en la elaboración del examen y las pruebas.
- Grado de interactividad y eficacia del interfaz creado.

Estas variables se ven complementadas por las variables de carácter específico relacionadas con la validación del proceso de realización del examen en lengua inglesa:

- La validación del proceso de realización del examen de lengua inglesa por parte de estudiantes con diferentes niveles de inglés.
- La validación del proceso de realización del examen de la lengua inglesa por parte de estudiantes con conocimientos mínimos de inglés.
- Memoria de recuerdo enfocado en la realización del examen en sí y basado en el nivel de conocimientos en lengua inglesa como uso de

vocabulario, uso de estructuras semánticas apropiadas, uso de preposiciones, pronombres, artículos, etc.

- Comprensión de los textos explicativos sobre las pruebas a realizar en la lengua inglesa
- Manejo de elementos multimedia en lengua inglesa.

Las preguntas creadas (gráficos 4 y 5) se enfocaron teniendo en cuenta las apreciaciones que previamente habían sido analizadas por los expertos mediante el método heurístico. De esta manera se pretendía confirmar y afianzar los puntos de vista que en algún caso se habían mostrado críticos.

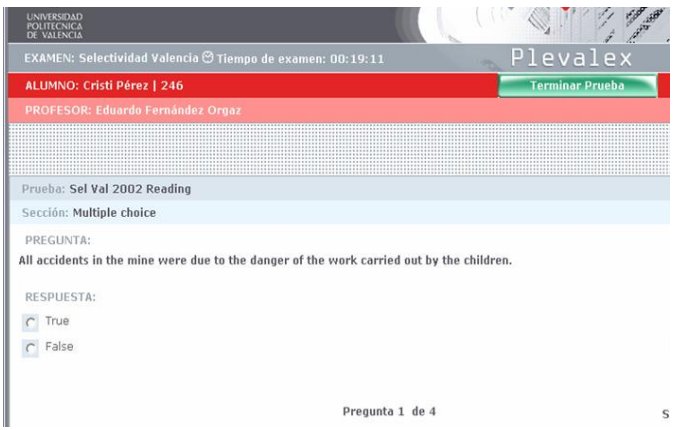

Gráfico 4. Visualización de una prueba de selección simple de un examen en la aplicación de *Frontoffice*.

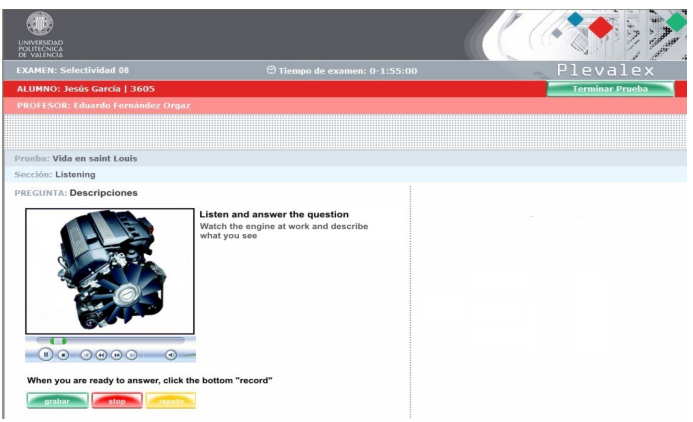

Gráfico 5. Visualización de una prueba con video de un examen en la aplicación *Frontoffice*.

Por ejemplo, en el segundo criterio heurístico planteado, concordancia entre el lenguaje del usuario y el mundo real, se indicaba que era necesario comprobar si el lenguaje era el adecuado para la realización de exámenes de la lengua inglesa debido tanto al carácter específico de las pruebas a nivel internacional como a los requerimientos de las distintas universidades.

5. Resultados cualitativos y conclusiones

A nivel global las respuestas fueron satisfactorias y, gracias a ello, la Universidad Politécnica de Valencia ha ido introduciendo y mejorando ciertos aspectos que parecían más débiles, como la calidad de imagen o la capacidad de ciertos estudiantes de moverse entre las distintas tareas presentadas en la plataforma. También se observó, tal y como en su momento había anticipado Fulcher (2003), que los colores y los iconos deben ser esmeradamente cuidados. Por ejemplo, en el diseño inicial se usaban botones verdes y rojos, ignorando involuntariamente que ese detalle podía afectar de manera relevante a la realización del examen por ciertos alumnos (por ejemplo, los que sufriesen daltonismo). Sin embargo, el resultado fundamental es precisamente la propia operatividad de la plataforma que, aún en su fase experimental, ya es seguida de cerca por empresas comerciales, institutos de lenguas y centros de investigación.

La plataforma PLEVALEX se encuentra en un estadio muy avanzado de validación. Basándonos en la experiencia realizada y analizando los resultados hasta el momento, podemos decir que es necesario establecer criterios mixtos o combinados a la hora de valorar otras plataformas como PLEVALEX: de sujetos implicados en el proceso de los exámenes mismos (como alumnos y profesores) y de expertos informáticos. En el caso de PLEVALEX se pudo asimismo contar con alumnos de inglés que estudian Comunicación Audiovisual y que fueron capaces de evaluar y comentar el aspecto educativo y el diseño formal de la plataforma.

En lo referente al procedimiento, al aplicar de forma complementaria los tres métodos ya expuestos, se han obtenido valoraciones expertas (informes bajo criterios heurísticos planteados por Nielsen y Molich, 1990). También se han obtenido valoraciones a nivel de usuario experto e inexperto en el conocimiento de la interactividad y multimedia que han ayudado a la puesta a punto de la plataforma y a su implementación final como una herramienta fiable y efectiva.

Por tanto, como se ha tratado de mostrar en este artículo, su validación se ha realizado desde un punto de vista didáctico, lingüístico y técnico. De cara al futuro, se espera poder utilizar este tipo de plataformas para la realización de exámenes *high stakes* como la Prueba de Acceso a la Universidad. En este sentido, es necesario seguir trabajando en el diseño de este tipo de herramientas en Internet y en la búsqueda de mejores exámenes de evaluación. En conclusión, PLEVALEX podría abrir una línea de cambio en el panorama de la evaluación asistida por ordenador y, muy especialmente, en el contexto del inglés como lengua de especialidad.

Agradecimientos

PLEVALEX se compone de las herramientas HIEO e HIELE. La Herramienta Informática de Evaluación Oral (HIEO) no se podría haber elaborado de no ser por la subvención recibida de la Generalitat Valenciana (Proyecto GV043/436). Por su parte, HIELE se ha podido desarrollar gracias al proyecto concedido por la Universidad Politécnica de Valencia (Proyecto 20040941).

(Artículo revisado recibido en junio de 2007)

Bibliografía

Alderson, J.C. y L. Hamp-Lyons (1996). "TOEFL preparation courses: A study of washback". *TESOL Quarterly* 13: 280-297.

Alderson, J.C. y A.H. Urquhart (1985). "The effect of students' academic discipline on their performance on ESP reading tests". *Language Testing* 2: 192-204.

Anglin, G.J., H. Vaez y K.L. Cunningham (2002). *Visual Representations and Learning: The Role of Static and Animated Graphics.* Lexington: Kentucky University Press.

Alvermann, D.E. y C.R. Hynd (1989). "Effects of prior knowledge activation modes and text structure on non-science majors' comprehension of physics". *Journal of Education Research* 83: 97-102.

Cargill, M. (2004). "Transferable skills within research degrees: a collaborative genre-based approach to developing publication skills and its implications for research education". *Teaching in Higher Education* 9: 83-98.

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

Chen, H.C. y M.F. Graves (1995). "Effects of previewing and providing background knowledge on Taiwanese college students' comprehension of American short stories". *TESOL Quarterly* 29: 663-686.

Connelly, M. (1997). "Using c-tests in English with post-graduate students". *English for Specific Purposes* 16: 139-150.

Cumming, A., L. Grant, P. Mulcahy-Ernt y D. Powers (2005). *A teacher-verification study of prototype speaking and writing tasks for new TOEFL. (TOEFL Monograph Series Rep. No. 26)*. URL: http://ftp.ets.org/pub/toefl/998777.pdf [28/06/07].

Douglas, D. y J. Smith (1997). *Theoretical underpinnings of the Test of Spoken English revision project. (TOEFL Monograph Series Rep. No. 9)*. Princeton, NJ: Educational Testing Service. URL: http://www.ets.org/Me-

dia/Research/pdf/RM-97-02.pdf [28/06/07].

Folse, K. (2006). *The Art of Teaching Speaking*. Michigan: Michigan University Press.

Fulcher, G. (1999). "Assessment in English for academic purposes: putting content validity in its place". *Applied Linguistics* 20: 221-236.

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

Fulcher, G. y R. Márquez Reiter (2003). "Task difficulty in speaking tests". *Language Testing* 20: 321-344.

García Laborda, J. (2004). "HIEO: Investigación y desarrollo de una herramienta informática de evaluación oral multilingüe". *Didáctica, Lengua y Literatura* 16: 77-88.

García Laborda, J. (2006a). "PLEVALEX: a new platform for oral testing in Spanish". *Eurocall Review* 9: 4-7. URL: http://www.eurocall-languages.org/news/newsletter/9/index.html [28/06/07].

García Laborda, J. (2006b). "¿Qué pueden aportar las nuevas tecnologías al examen de selectividad de inglés? Un análisis de fortalezas y oportunidades". *Revista de CC Educación* 201: 151-166.

García Laborda, J. y L.G. Bejarano (2005). "Análisis de la necesidad de creación de páginas web para la evaluación y baremación de estudiantes internacionales: una experiencia internacional" en M.L. Carrió Pastor (ed.), *Perspectivas interdisciplinares de la lingüística aplicada*, 399-404. Valencia: Universidad Politécnica de Valencia.

García Laborda, J. y E. Enríquez Carrasco (2005). "¿Es HERMEX una plataforma válida para diagnosticar lingüísticamente? Un análisis funcional". *TEAM,* 3: 30-32.

Hamp-Lyons, L. (1998). "Ethical test preparation practice: the case of the TOEFL". *TESOL Quarterly* 32: 329-337.

Herrera Soler, H. (2005). "El test de elección múltiple: herramienta básica en la Selectividad" en Herrera Soler, H. y J. García Laborda (eds.), *Estudios y criterios para una selectividad de calidad en el examen de inglés*, 65-98. Valencia: Universidad Politécnica de Valencia.

Jackson, J. (2005). "An inter-university, cross-disciplinary analysis of business education: perceptions of business faculty in Hong Kong". *English for Specific Purposes* 24: 293-306.

Jacoby, S. y T. McNamara (1999). "Locating competence". *English for Specific Purposes* 18: 213-241.

Jamieson, J., S. Jones, I. Kirsch, P. Mosenthal y C. Taylor (2000). *TOEFL 2000 Framework: A Working Paper (TOEFL Monograph Series Rep. No. 16)*. Princeton, NJ: Educational Testing Service URL: http://www.ets.org/Media/Research/pdf/RM-00-06.pdf [28/06/07].

Jeffries, R., J.R. Millar, C. Wharton y K.M. Uyeda (1991). "User interface evaluation in the real world: A comparison of four techniques". *Proceedings ACM CHI´91 Conference*, 119-124. Seattle: ACM Press.

Jensen, C. y C. Hansen (1995). "The effect of prior knowledge on EAP listening-test performance". *Language Testing* 12: 99-119.

Krekeler, C. (2006). "Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited". *Language Testing* 23: 99-130.

Kobrin, J.L. y J.W. Young (2003). "The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration". *Applied Measurement in Education* 16: 115-140.

Lewis, C.P., C. Wharton y J. Rieman (1990). "Testing a walkthrough methodology for theory-based design of walk-up and use-interface". *Proceedings ACM CHI´90 Conference*, 235-242. Seattle: ACM Press.

Lewkowicz, J.A. (2000). "Authenticity in language testing: some outstanding questions". *Language Testing* 17: 43-64.

Nielsen, J. (1994). "Enhancing the explanatory power of usability heuristic". *Proceedings ACM CHI´94 Conference*, 152-158. Seattle: ACM Press.

Nielsen, J. y R.L. Mack (1994). *Heuristic Evaluation. Usability Inspection Methods*. New York: John Wiley and Sons.

Nielsen, J. y R. Molich (1990). "Heuristic evaluation of user interfaces". *Proceedings ACM CHI´90 Conference*, 249-256. Seattle: ACM Press.

Oller, J.W. Jr. (1979). "Explaining the reliable variance in tests: the validation problem" en E.J. Briere y F.B. Hinofotis (eds.), *Concepts in Language Testing: Some Recent Studies*, 61-74. Washington, D.C.: TESOL.

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

Roever, C. (2001). "Web-based language testing". *Language Learning & Technology* 5: 84-94.

Salmani-Nodoushan, M.A. (2003). "Text familiarity, reading tasks, and ESP test performance: a study on Iranian LEP and Non-LEP university students". *Reading Matrix: An International Online Journal* 3: 1-14. URL: http://www.readingmatrix.com/articles/nodoushan/article.pdf [28/06/07].

Spence-Brown, R. (2001). "The

eye of the beholder: authenticity in an embedded assessment task Language Testing". *Language Testing* 18: 463-474.

Yoshida, K. (1998). *Student Recommendations for ESP Curriculum Design.* Ann Arbour, MI: ERIC Reproduction Service. Documento ED424782. URL: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/17/07/08.pdf [28/06/07].

Wise, S.L. y B.S. Plake (1989). "Research on the effects of administering tests via computers". *Educational Measurement: Issues & Practice* 8: 5-10.

Zareva, A. (2005). "What is new in the new TOEFL-iBT 2006 test format?". *Electronic Journal of Foreign Language Teaching* 2: 45-57. URL: http://e-flt.nus.edu.sg/v2n22005/zareva.htm [28/06/07].

Dr. Jesús García Laborda es doctor en Filología Inglesa por la Universidad Complutense de Madrid. Imparte clases de inglés para Turismo y Telecomunicaciones en la Universidad Politécnica de Valencia. Ha realizado trabajos de investigación, comunicaciones, rescensiones y publicaciones en exámenes de bajo impacto asistidos por ordenador.

Dra. Teresa Magal Royo es doctora en Bellas Artes por la Universidad Politécnica de Valencia y especialista en diseño gráfico de software. Imparte clases en Comunicación Audiovisual en la misma universidad y ha realizado trabajos de investigación, comunicaciones y publicaciones en la misma materia, software educativo y diseño de plataformas digitales.

NOTAS

1 La empresa TSE que distribuye TOEFL plantea varios tipos de tareas más concretas que, en realidad, se resumen en las anteriormente citadas. Consúltese la dirección URL: http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=f0dc01c203d95010VgnVCM10000022f95190RCRD&vgnextchannel=ab16197a484f4010VgnVCM10000022f95190RCRD) [28/06/07].

2 *Dialang* es una herramienta de evaluación en línea desarrollada con el apoyo de la Unión Europea y basada en el Marco de referencia común europeo. Esta plataforma ofrece pruebas diagnósticas en catorce lenguas y está disponible de forma gratuita en la dirección electrónica http://www.dialang.org.