La certificación del español con fines profesionales. El examen adaptativo LanguageCert USAL esPro. Juan Miguel Prieto Hernández. Universidad de Salamanca

Juan Miguel Prieto Hernandez

1. Introducción

La enseñanza del español con fines específicos surge a finales de los años 60 del siglo pasado de forma casi simultánea a la aparición del enfoque comunicativo en el aprendizaje de lenguas extranjeras. La denominación EFE (*Español con fines específicos*) se generaliza a comienzos de la década de los años 90 y, desde aquel momento, se convierte en una de las ramas más relevantes de la enseñanza del español como lengua extranjera. La evolución que ha seguido este campo en las instituciones educativas públicas y privadas especializadas ha llevado a establecer una división en dos grandes grupos: el *Español con fines profesionales* (EFP) y el *Español con fines académicos* (EFA).

La enseñanza del español con fines profesionales se centra en los procesos de enseñanza-aprendizaje de la lengua especializada que utilizan los profesionales que trabajan en diversos contextos laborales. Nos referimos al mundo laboral en el que previsiblemente ingresarán los estudiantes que finalicen la Enseñanza Secundaria Obligatoria (ESO), la Formación profesional, el Bachillerato y los diferentes grados universitarios.

En 1997 la Universidad de Cambridge, la Alianza francesa, el Instituto Goethe y la Universidad de Salamanca diseñaron el sistema de certificación BULATS con el propósito de medir el nivel de competencia y dominio del español con fines profesionales. Se trataba de un sistema de exámenes centrado en el ámbito laboral que ayudaba a las empresas a tomar decisiones relacionadas con la formación, contratación, promoción o movilidad de los trabajadores, a la vez que estos podían acreditar su nivel. Se desarrolló en cuatro idiomas (inglés, francés, alemán y español) y permitía determinar de forma rápida y fiable (Geranpayeh 2001a) el nivel de dominio lingüístico de cualquier usuario en relación con la

escala de seis niveles del *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación* (Consejo de Europa 2002) (en adelante MCER).
El primer BULATS era un examen progresivo que se realizaba en papel. En el año 2000 se implementó la modalidad en ordenador y, en 2008, se lanzó al mercado la primera versión adaptativa de una prueba de Comprensión auditiva y de lectura con fines profesionales. Se trataba de un examen que se ajustaba al nivel que el candidato iba acreditando que tenía, de forma que, según la respuesta que este diera a cada pregunta, el sistema seleccionaba tareas más fáciles o más difíciles con el fin de no ofrecerle preguntas excesivamente sencillas o complicadas. La calidad la certificación BULATS en español se vio reconocida en 2012, fecha en la que la plataforma de certificación universitaria CertiUni, promovida ese mismo año por la conferencia de rectores de las universidades españolas (CRUE), seleccionó BULATS para que los universitarios españoles pudieran acreditar su nivel de conocimientos lingüísticos. De este modo, Cursos Internacionales de la Universidad de Salamanca se convirtió en el único suministrador de los exámenes que se desarrollaran a través de este medio en las universidades españolas.

En 2017 el estudio salmantino comenzó a desarrollar en solitario la versión en español de BULATS. Para ello se asoció con LanguageCert, empresa filial de PeopleCert, para seguir ofreciendo en todo el mundo, a través de su plataforma online, la versión renovada del examen, llamado ahora Languagecert USAL esPro. Esta alianza supuso un gran avance en el campo de la evaluación lingüística, ya que combinaba la experiencia de la Universidad de Salamanca, líder en la enseñanza y evaluación del español como lengua extranjera, con la capacidad de innovación tecnológica de PeopleCert en el ámbito de la certificación, del desarrollo de exámenes y de su administración.

2. Desarrollo de un sistema de certificación

En el desarrollo de BULATS se siguieron los parámetros que actualmente están consensuados por la comunidad científica para el diseño de sistemas de certificación. Estas

prácticas, seguidas hoy en día por los sistemas de certificación más prestigiosos, resultaron pioneras en los años 90 y establecieron unos requisitos de calidad y rigor que marcaron un punto de inflexión en las exigencias de calidad de los sistemas de certificación de lenguas.

2.1. Estudio de necesidades

Primeramente se realizó un estudio de necesidades con el fin de analizar la necesidad de desarrollar un sistema de certificación del idioma español con fines profesionales. Los resultados avalaron esa conveniencia y se diseñó un sistema de exámenes para acreditar el grado de competencia del idioma español en el ámbito profesional, uno de los cuatro ámbitos que describe el MCER, junto con el educativo, el público y el profesional.

2.2. Delimitación del marco conceptual

Una vez tomada la decisión de desarrollar el sistema de certificación se comenzó a trabajar en el diseño de la prueba con una delimitación del marco conceptual, que incluía la descripción del constructo que debía evaluarse, las características del modelo de lengua que se iba a utilizar (ámbito profesional), las consecuencias de la evaluación y las inferencias que podrían realizarse a partir de los resultados de esta.

2.3. Redaccion de especificaciones

Seguidamente comenzó el proceso de redacción de las especificaciones del examen. En ellas se facilita información sobre lo que el examen evalúa y cómo lo evalúa, figuran los contenidos de los procesos cognitivos, las características psicométricas de las tareas, así como otras informaciones pertinentes para el proceso evaluativo (Lane y Stone 2006: 390) como, por ejemplo, el plan que deben seguir los redactores de las tareas de examen. En las especificaciones también se informa acerca de los procedimientos de administración de la prueba: instrucciones, duración, la manera de puntuar, criterios de calificación, etc., y se definen las características de la población meta y la previsión de las inferencias de los

resultados. Igualmente se precisa lo que se debe tener en cuenta en relación con la validez, la confiabilidad, la ética y la justicia y la planificación del trabajo.

2.4. Diseño de las tareas

Posteriormente se pasó al diseño de las tareas y de las pruebas del examen, a su desarrollo y al proceso de experimentación y pilotaje. Las tareas que constituyen una prueba de desempeño deben representar de manera sistemática la intención del constructo, han de ser válidas, fiables y equitativas, además de estar estandarizadas. Cada tarea que integre la prueba debe servir para medir algo único y diferente respecto a las otras tareas y tiene que garantizar la comparabilidad de resultados en el tiempo (Haertel y Linn 1996 y American Educational Research association, American Psychological Association y National Council on Measurement in Education 1999). Asimismo, se detalló el sistema de evaluación y los criterios y escalas de calificación de las pruebas de Expresión oral y Expresión escrita. Finalmente se procedió a redactar toda la documentación necesaria para las diversas partes interesadas: candidatos, creadores de pruebas, empresas, etc.

Tanto el proceso de elaboración de las especificaciones de la prueba como el del diseño de las tareas es recurrente y, en consecuencia, cualquier problema o disfunción detectado durante el proceso de elaboración del examen puede ocasionar que haya que modificar las especificaciones y reemprender el proceso. No hay que perder de vista que objetivo final es que el examen construido sea fiable y válido, y que las conclusiones que se extraigan de sus resultados sean significativas, útiles y apropiadas.

2.5. Escala de niveles

La escala de niveles que se utilizó como referencia es la que describe el MCER en el que se establece una serie ascendente de niveles comunes para describir el dominio que los candidatos tienen de la lengua (del A1 al C2). Como BULATS se concibió algunos años antes de la aparición del MCER, se utilizó la escala de cinco niveles de ALTE que, en términos generales, se corresponden con los del MCER (tabla 1).

Niveles del MCER	A1	A2	B1	B2	C1	C2
Niveles de ALTE	Nivel Acceso de ALTE	Nivel 1 de ALTE	Nivel 2 de ALTE	Nivel 3 de ALTE	Nivel 4 de ALTE	Nivel 5 de ALTE

Tabla 1. Relación entre la estructura de niveles de ALTE y la escala de niveles del MCER

Posteriormente, para vincular las tareas y los exámenes con la escala de niveles del MCER se siguió uno de los procedimientos descritos en la publicación del Consejo de Europa *Relating Language Examination to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A manual*. Se comenzó con una recogida de datos para relacionar las autoevaluaciones de los cuestionarios «Puede hacer» de ALTE con las calificaciones obtenidas por los candidatos en el examen en los distintos niveles y se constató que existía «una relación muy clara que hizo posible comenzar a describir el sentido de una calificación de examen en función de perfiles característicos de capacidad lingüística ”Puede hacer”» (MCER 238). Seguidamente se recogieron respuestas en las que el anclaje se lograba a partir de la redacción preliminar de los cuestionarios detallados del MCER que presentaron en 1996 John Leslie Melville Trim (Director del proyecto), Brian North y Daniel Coste. Los resultados de la validación psicométrica de los datos demostraron la existencia de una excelente correlación entre los cuestionarios «Puede hacer» que se habían utilizado en el proceso de vinculación de BULATS y las escalas ilustrativas del MCER (MCER 239). Los ítems que configuran las diversas pruebas de la versión adaptativa del examen se seleccionan automáticamente para cada usuario dependiendo de su nivel y se vinculan a la escala de niveles del MCER mediante el modelo de Rasch (Geranpayeh 2001b, 2001c). Recordemos que, cuando se desarrolló la versión adaptativa de BULATS en 2008, ya hacía seis años que había sido publicado el MCER.

En el certificado que se expide al finalizar el examen se informa al usuario del nivel alcanzado en una escala de 0 a 100, en la que el 0 se situaría en el nivel más bajo de la escala y 100 en el más alto. Esta escala es una modificación o aclaración de los valores escalares de los candidatos y de los ítems, que en el modelo de Rasch habitualmente se expresan en la escala *logit*. La escala se subdivide a su vez en seis bloques que se corresponden con los niveles de la escala de ALTE y del MCER. También se informa al

usuario de lo que puede ser capaz de hacer lingüísticamente en el idioma español, según el nivel que haya alcanzado.

2.6. Administración del test

Finalmente se detallaron los objetivos y procesos de la administración del test. También se tuvo en cuenta el formato de los documentos en los que se presentarían los resultados de los análisis realizados y la monitorización de las escalas de revisión.

La versión en papel de la prueba de Comprensión auditiva y de lectura de USAL esPro es progresiva, lo que significa que la dificultad de las tareas se va incrementando a medida que transcurre el examen. La versión en ordenador que se implementó en el año 2000 tenía las mismas características que la versión en papel (progresividad), salvo el tipo de soporte utilizado. Para la versión adaptativa de USAL esPro se han empleado ítems experimentados y calibrados para el sistema de certificación BULATS y se ha comprobado que los resultados obtenidos en las versiones progresiva y adaptativa del examen son equivalentes.

3. Características de una prueba adaptativa

Comentábamos arriba que los exámenes adaptativos se generan dinámicamente basándose en las respuestas de los candidatos: en el caso de que fallen, la siguiente tarea será más sencilla y, si aciertan, se les propondrá otra más difícil. El objetivo es ajustarse en la medida de lo posible al nivel de competencia del candidato puesto que, cuanto mayor sea este ajuste, mayor información útil se obtiene (Green, Bock et al. 1984).

3.1. Plataforma informática

Para el desarrollo de una versión adaptativa es necesario disponer de una plataforma informática (banco de tareas) que permita la gestión integral de las tareas y de los ítems de

las pruebas desde el momento de su creación hasta la administración del examen. El programa debe ser capaz de predecir, a partir de las respuestas dadas por el candidato, cómo respondería este a cualquiera de los ítems que aún no se le han presentado, de seleccionar y ofrecerle a continuación la tarea más apropiada y de dar, al final, una puntuación que represente la habilidad del examinado (Lord, 1974).

Las tareas deben estar almacenadas de manera estructurada junto con sus características psicométricas y de contenido. La información archivada debe incluir, para cada ítem, un identificador único, el enunciado y los materiales asociados (tarea a la que corresponde, tablas, fotos, audios, etc.), la opción correcta y los distractores, una referencia al dominio específico que evalúa, el número de veces que ha sido administrado, y los parámetros del modelo de la TRI relativos tanto al ítem como a la tarea en la que se integra.

Para conocer la información psicométrica asociada al ítem, todas las tareas de la plataforma informática deben haber sido experimentadas, como mínimo, con 200 estudiantes. También es conveniente que entre los alumnos con los que se experimentan las tareas haya un equilibrio entre hombres y mujeres, y que las nacionalidades de los alumnos con los que se realiza la experimentación sean similares a las de los candidatos que habitualmente se inscriben en el examen.

3.2. Análisis psicométrico

En el análisis psicométrico de las versiones preliminares de las pruebas de Comprensión auditiva y de lectura que se realiza tras la experimentación se integran estadísticos procedentes de dos modelos de medida: el modelo lineal clásico (Teoría Clásica de los Tests, TCT) y un modelo de la Teoría de Respuesta al Ítem (TRI): el Modelo de Rasch (1960/1980). En la Teoría Clásica de los Tests (TCT) se interpreta que la puntuación en un examen es la suma de la puntuación verdadera obtenida por el candidato en la prueba y de un error no sistemático de medida. Para determinar cuál es la proporción de la varianza verdadera en relación con la varianza observada, es decir, cuál es la proporción de las diferencias entre las puntuaciones del examen que no se debe a los errores de medida, en la

TCT su emplea habitualmente el *coeficiente de fiabilidad*, que oscila entre 0 y 1. Un procedimiento muy utilizado para estimarlo es el coeficiente alfa de Cronbach (Abad et al. 2011). Según la Federación Europea de Asociaciones de Psicólogos (EFPA 2013) son inadecuados los ítems con valores inferiores a .70, adecuados los coeficientes entre .70 y .80, buenos las magnitudes que oscilan entre .80 y .90, y excelentes los valores superiores a .90 (Prieto 2014).

Además, en el análisis realizado por medio de la TCT se tiene en cuenta, por un lado, la dificultad de los ítems, es decir, la proporción de alumnos que los acierta y, por otro, su discriminación, esto es, la correlación entre el ítem y la puntuación total de la prueba. También se suele tener en cuenta si los alumnos que seleccionan la opción correcta tienen mayor puntuación promedio en el test que los que eligen las incorrectas (coherencia) y si las opciones incorrectas atraen a un porcentaje suficiente de alumnos (eficacia). Puede consultarse a este respecto Gulliksen (1950), Martínez, Hernández y Hernández (2006) y Muñiz (2000).

A pesar de sus evidentes ventajas, de su sencillez matemática y de su enjundia psicológica (Muñiz 2000), hay cuestiones que no es posible resolver mediante la TCT. Su principal limitación radica en que los índices psicométricos de la dificultad y de la discriminación de los ítems están en estrecha relación con el nivel de competencia de los alumnos que han participado en la experimentación y, a la vez, que sus puntuaciones dependen de las características de los ítems incluidos en la prueba. Esta dependencia dificulta la comparación entre alumnos que han realizado tests diferentes y la construcción de bancos de ítems calibrados en la misma métrica. Todo ello hace aconsejable utilizar modelos de medida derivados de la Teoría de Respuesta al Ítem (TRI), entre los que destaca el modelo de Rasch, dado que permite medir de forma conjunta a alumnos e ítems en una escala común (mapa de la variable), cuantificar el error típico de medida de las puntuaciones, y determinar de manera objetiva el nivel de competencia de los alumnos y la dificultad de los ítems.

Resulta relativamente sencillo detectar los ítems y las personas que no se ajustan al modelo por medio de los estadísticos de ajuste de datos *infit* y *outfit*, en cuanto que permiten cuantificar la magnitud de los *residuos* (diferencias entre las respuestas observadas y las esperadas). También se puede saber cuál es la precisión con que las medidas diferencian a las personas en el atributo que se evalúa mediante el estadístico denominado *Person Separation Reliability* (PSR) (Eckes, 2011). De forma similar se interpreta *Item Separation Reliability* (ISR), que es el cociente entre la varianza de las medidas verdaderas de los ítems en la variable latente y la varianza observada. Por medio de la aplicación del modelo de Rasch es posible calcular, igualmente, la precisión de las estimaciones de la medida de cada alumno y de la dificultad de cada ítem mediante el error estándar de la persona o del ítem. Pero a diferencia de la TCT, en el Modelo de Rasch este error puede variar en su magnitud a lo largo de la variable (cuanto más difícil sea la prueba para los alumnos, menor será la precisión con la que se mida a los menos competentes y, cuanto más fácil sea, con menor precisión se medirá a aquellos con mayor nivel de competencia). Pueden consultarse excelentes descripciones del modelo de Rasch en Engelhard (2013) y Prieto y Delgado (2003).

El número de tareas y de ítems que requiere el uso de la tecnología adaptativa es elevado y para analizar el funcionamiento diferencial de los ítems (DIF) es necesario disponer de un número equilibrado de hombres y de mujeres en la experimentación con el fin de conocer si los ítems se comportan de manera diferente en los dos grupos. El DIF indica una diferencia del funcionamiento del ítem (o prueba) entre grupos de candidatos igualmente capaces que tienen una probabilidad distinta de responderlo con éxito (Potenza y Dorans 1995 y Prieto y Nieto 2014).

4. Ventajas de la tecnología adaptativa

Los sistemas de certificación con tecnología adaptativa presentan evidentes ventajas frente a las pruebas progresivas: se reduce el tiempo de aplicación (la prueba adaptativa de USAL

esPro dura casi la mitad que la versión progresiva en papel); se incrementa la seguridad y se reduce la posibilidad de copia y de trampa; se incrementa el control sobre la administración de la prueba; se mejora el procesamiento de las respuestas y su interpretación; se realizan estimaciones más precisas que en las pruebas progresivas, especialmente en candidatos de poca o mucha habilidad, ya que todas las tareas a las que responden tienen un índice de dificultad relacionado con su nivel de competencia; y permite, además, que la información de los resultados de la prueba se dé en mucho menos tiempo que en las versiones en papel.

5. Bibliografía

ABAD, F. J.; OLEA, J.; PONSODA, V. y GARCÍA, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION Y NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

CONSEJO DE EUROPA (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Madrid: MECD y Anaya. En: <http://cvc.cervantes.es/obref/marco> Fecha de consulta 14.8.2019.

CONSEJO DE EUROPA (2009). *Relating Language Examination to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A manual*. Estrasburgo: Consejo de Europa. En: <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d>. Fecha de consulta 14.8.2019.

CONSEJO DE EUROPA (2011). *Manual for Language Test Development and Examining for Use with the CEFR. Produced by ALTE on Behalf of the Language Policy Division, Council of Europe*. Estrasburgo: Consejo de Europa. En: <https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b> Fecha de consulta 14.8.2019.

ECKES, T. (2011). *Introduction to Many-facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt: Peter Lang.

EFPA (2013). *EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests* (Version 4.2.6). European Federation of Psychologists' Associations. En: <http://www.efpa.eu/professional-development> Fecha de consulta 14.8.2019.

ENGELHARD, G. (2013). *Invariant Measurement. Using Rasch Models in the Social, Behavioral, and Health Sciences*. Londres: Routledge.

ENSEÑANZA de la lengua para fines específicos. En AA.VV. (2008). *Diccionario de términos clave de ELE*. Madrid: SGEL; Instituto Cervantes.

GERANPAYEH, A. (2001a). CB BULATS: «Examining the reliability of a computer based test using test- retest method». *Research Notes* 5, 14-16.

GERANPAYEH, A. (2001b). *Linguaskill Annual Report 2001*. Cambridge ESOL Internal (Publicación de uso interno).

GERANPAYEH, A. (2001c). *Linguaskill English 2001*. Cambridge ESOL. (Publicación de uso interno).

GREEN, B. F., Bock, R. D., HUMPHREYS, L. G., LINN, R. L. y RECKASE, M. D. (1984). «Technical guidelines for assessing computerized adaptive tests». *Journal of Educational Measurement* 21(4): 347-360.

GULLIKSEN, H. (1950). *Theory of mental tests*. Nueva York: Wiley.

HAERTEL, E. H. y LINN, R. L. (1996). «Comparability», en: G. W. Phillips (ed.), *Technical Issues in Large-scale Performance Assessment* (NCES 96-802). Washington, DC: National Center for Education Statistics, 59-78.

LANE, S. y STONE, C. A. (2006). «Performance assessment», en: R. L. Brennan (ed.). *Educational Measurement*, (4ª edición). Westport, CT: American Council on Education and Praeger, 387–431.

LORD, F. M. (1974). «Individualized testing and item characteristics curve theory». Capítulo en *Contemporary developments in mathematical psychology*, Vol. II. Eds. D. H. Krantz, R. C. Atkinson, R. D. Luce, y P. Suppes. San Francisco, California (USA): Freeman.

MÁRTÍNEZ, M. R.; HERNÁNDEZ, M. J. y HERNÁNDEZ, M. V. (2006). *Psicometría*. Madrid: Alianza Editorial.

MUÑIZ, J. (2000). *Teoría Clásica de los Tests*. Madrid: Pirámide.

POTENZA, M. T., y DORANS, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23-37.

PRIETO, G. y DELGADO, A. R. (2003). «Análisis de un test mediante el modelo de Rasch», *Phicothema*, 15, 1, 94-100.

PRIETO, G. (2014). *Breve descripción de la metodología psicométrica empleada en el análisis de los exámenes del DELE*. Instituto Cervantes – Universidad de Salamanca. (Informe interno de marzo de 2014).

PRIETO, G. y NIETO, E. (2014). Influence of DIF on Differences in Performance of Italian and Asian Individuals on a Reading Comprehension Test of Spanish as a Foreign Language. *Journal of Applied Measurement, 15*, 176-188.

RASCH, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhague: Danish Institute for Educational Research, 1960 / Chicago, IL: University of Chicago Press, 1980.