DATOS DE VALIDACIÓN Y FIABILIDAD DEL EXAMEN

Convocatoria de junio de 2014 (examen de muestra en la web)

Fiabilidad del examen	.909
Fiabilidad de la CA	.883
Fiabilidad de la CL	.752
Discriminación de los ítems	Se utilizaron aquellos entre .250 y .643
Nivel de los ítems	Se utilizaron las escalas *logit"* (para B1 entre -1.23 y 0.71; para B2 entre 0.72 y 2.79) (North y Jones, 2009)
Correlación entre las pruebas de respuesta preseleccionada (CA y CL)	0.6
Establecimiento de puntos de corte en la CA y CL	Sistema de tres rondas (expertos, expertos con datos de impacto, expertos con datos de IRT)
Porcentaje de aciertos para cada nivel en la CA y CL	Cada nivel tiene un número mínimo de ítems que puede variar ligeramente en cada convocatoria. Para superar cada prueba se deben responder correctamente los ítems que se han considerado mínimos de cada uno de los niveles mediante un proceso de calibración con expertos y análisis psicométricos.
Fiabilidad de la calificación de la EIE y la EIO	Tanto la evaluación de la EIE como la de la EIO se realizan por el sistema de doble corrección. En la EIO uno de los evaluadores tiene el papel de entrevistador. Los evaluadores llevan a cabo sesiones periódicas de estandarización de criterios.

A continuación se explican de forma detallada los resultados más relevantes del proceso de análisis de validación llevado a cabo con las pruebas, sus ítems y el examen en conjunto, con el fin de elaborarlo de forma adecuada a los niveles que pretende medir y a sus objetivos. En nuestro examen hacemos los análisis utilizando SPPS (Statistical Package for the Social Sciences), de IBM, y un análisis de IRT (Teoría de respuesta al ítem). Con ellos, además de la fiabilidad, obtenemos datos como medias, discriminación, correlación entre pruebas, nivel de cada ítem, índice de dificultad de cada ítem con independencia de la población con la que se haya probado, errores de medida del examen en su conjunto y de cada uno de los ítems, número de niveles que es capaz de diferenciar, etc.

FIABILIDAD Y DISCRIMINACIÓN DE LOS ÍTEMS Y PRUEBAS

Fiabilidad

La fiabilidad es el valor con el que un *test* mide lo que pretende estar midiendo, con la seguridad de que en la obtención de la nota no influyan otros elementos ajenos a la habilidad lingüística que se quiere calibrar. Ello nos asegura que lo que evalúa se mantiene estable y solo afectado por incrementos de la habilidad lingüística. La fiabilidad se obtiene mediante programas informáticos y su resultado es siempre entre -1 y +1, siendo +1 la mejor medida, lo que supone que si el *test* se repitiera dos veces con los mismos candidatos, obtendrían exactamente los mismos resultados si estos no hubieran aprendido nada nuevo. Las cifras aceptables para asegurar que una prueba tiene una buena fiabilidad no deben estar por debajo de .7 (0.7), siendo mejores, como hemos dicho, cuanto más próximos están al 1.

En la primera convocatoria de nuestro examen, la fiabilidad total fue de .909 y la de sus pruebas de comprensión auditiva y lectora, .883 y .752, respectivamente.

Discriminación

La discriminación es la capacidad que posee un ítem para diferenciar entre candidatos de más nivel de habilidad lingüística y candidatos de menor habilidad. Son ítems buenos aquellos que solamente se responden bien por los candidatos que poseen la habilidad y el nivel suficiente y, al contrario, se responden mal por los candidatos que no poseen suficiente nivel o habilidad. Esto no es así siempre, pues puede suceder que, por ambigüedad en el diseño de los ítems, por un mal diseño, o por otras razones, haya ítems que se con*test*en bien por candidatos que no poseen el nivel adecuado, o que los fallen los que, sin embargo, sí posea el nivel. En estos casos el ítem no sirve para decir quiénes son los que tienen mejores habilidades y los que no: el ítem no es bueno y debe desecharse. Una buena discriminación en SPSS está entre .250 y 1000, siendo mejor cuanto más alta sea (Green, 2013). Los ítems con los que finalmente se compuso nuestro examen se sitúan en unos índices de entre .250 y .643.

Con los resultados del pilotaje y análisis de ítems y tareas (fiablidad, discriminación, error de medida, etc.) se reforma y diseña finalmente el examen, seleccionando los ítems que tienen el nivel apropiado para conseguir un *test* binivel (B1 y B2) suficientemente significativo y con los que se puedan obtener datos con los que elaborar una calificación final justa.

ESTABLECIMIENTO DE LOS NIVELES QUE EVALÚA EL EXAMEN

Para asegurarnos de que el nivel de los ítems de las tareas finales del examen en su totalidad se adecua al nivel de dominio previsto, tanto de candidatos de B1 como de B2, utilizamos un análisis de IRT (Teoría de respuesta al ítem). Este método calcula, entre otras cosas, el nivel y el valor de error de medida de cada ítem con independencia de la población con la que se haya probado. Posteriormente, utilizamos el método de Angoff para determinar los puntos de corte (como se explica abajo).

Los datos obtenidos mediante IRT nos aportan información sobre cuestiones importantes para la composición del examen: el nivel de cada ítem (North y Jones, 2009), su margen de error de medida, su poder de discriminación, la cantidad de niveles que se pueden evaluar mediante esa batería de ítems, la uniformidad de la prueba (o su capacidad para evaluar las facetas deseadas, eliminando los ítems que se salgan del modelo general del test), etc.

Con estos datos y mediante el método Angoff ampliado a tres rondas (Figueras, 2011) se procede a una sesión de establecimiento de puntos de corte en la que un panel de expertos toma las decisiones y propone las puntuaciones con las que se superará el examen.

CORRELACIÓN ENTRE LAS DISTINTAS PRUEBAS

Una vez diseñadas las pruebas con los ítems apropiados, también comprobamos si la confección del examen sirve para evaluar realmente habilidades diferenciadas entre sí (CA, CL, EIE y EIO). Para esto hallamos sus coeficientes de correlación: la medida en que en dos pruebas hay elementos comunes que, por lo tanto, estarían siendo evaluados varias veces y pesarían más en la nota final del examen. Si lo que se desea diseñar son pruebas diferentes (una que evalúe la CA y otra la CL, p. e.), responder a estas exige poner en funcionamiento elementos cognitivos, estrategias y conocimientos diferentes. En tal caso la correlación debe ser baja, entre 0.4 y 0.6.

En nuestro caso, la correlación entre la CA y CL es de 0.6, lo que es prueba de que miden habilidades diferentes. Este método también se emplea para averiguar la relación entre cada una de la otras partes del examen (EIE, EIO, CA y CL) entre sí.

ESTABLECIMIENTO DE PUNTOS DE CORTE Y NOTA FINAL EN LAS PRUEBAS DE CA Y CL

Para asegurarnos de que el nivel de los ítems de las tareas finales del examen en su totalidad se adecua al nivel de dominio previsto, tanto de candidatos de B1 como de B2, utilizamos un análisis de IRT (Teoría de respuesta al ítem). Este método calcula, entre otras cosas, el nivel y el valor de error de medida de cada ítem con independencia de la población con la que se haya probado. Posteriormente, utilizamos el método de Angoff para determinar los puntos de corte (como se explica abajo).

Los datos obtenidos mediante IRT nos aportan información sobre cuestiones importantes para la composición del examen: el nivel de cada ítem (North y Jones, 2009), su margen de error de medida, su poder de discriminación, la cantidad de niveles que se pueden evaluar mediante esa batería de ítems, la uniformidad de la prueba (o su capacidad para evaluar las facetas deseadas, eliminando los ítems que se salgan del modelo general del test), etc.

Con estos datos y mediante el método Angoff ampliado a tres rondas (Figueras, 2011) se procede a una sesión de establecimiento de puntos de corte en la que un panel de expertos toma las decisiones y propone las puntuaciones con las que se superará el examen.

FIABILIDAD DE LA CALIFICACIÓN DE LA EIE Y LA EIO

En las pruebas de EIE y EIO, como se ha mencionado ya en el presente artículo, los candidatos son evaluados mediante una escala holística que ha sido desarrollada específicamente para el examen siguiendo criterios del MCER. La escala para EIE tiene en cuenta los siguientes criterios: cumplimiento de a la tarea, coherencia, alcance y dominio de vocabulario, y corrección. La escala para EIO tiene en cuenta los siguientes criterios: fluidez, coherencia, monólogo, corrección, alcance y dominio de vocabulario, e interacción. Tanto la evaluación de la EIE como la de la EIO se realizan por el sistema de doble corrección. En la EIO uno de los evaluadores tiene el papel de entrevistador. Los evaluadores llevan a cabo sesiones periódicas de estandarización de criterios.

REFERENCIAS BIBLIOGRÁFICAS:

CONSEJO DE EUROPA (2002), *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*, Madrid: Ministerio de Educación, Cultura y Deporte; Instituto Cervantes. [Título original: Common European Framework of Reference for Languages Learning, teaching, assessment, (2001), Strasbourg: Council ofEurope].

CONSEJO DE EUROPA (2009), *Relating Language Examinations to the CEF, Manual Preliminary Pilot Version. Introduction and Feedback to the pilot phase*, [pdf], Strasbourg: Council of Europe.

CONSEJO DE EUROPA (2011), *Manual for Language test Development and Examining*, [pdf], Strasbourg: Council of Europe.

FIGUERAS, N., F. KAFTANDJIEVA, S. TAKALA, (2011), Relating a Reading Comprehension Test to the CEFR Levels: A Case of Standard Setting in Practice with Focus on Judges and Items , The Canadian Modern Language Review/La Revue canadienne des langues vivantes, 69, 4, (November / novembre), 359 385.

GREEN, R. (2013), *Statistical Analyses for Language Testers*, Palgrave McMillan.

NORTH, B. / JONES, N. (2009), *Relating Language Examinations to the Common European* Framework *of Reference for Languages: Learning, Teaching, Assessment (CEFR).* Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling. [pdf], Strasbourg: Council of Europe, (consulta: 20 de febrero de 2014), <http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Publications/Standards_language_EN.pdf>

CIZEK, G. /BUNCH, M. (2007), *Standard Setting*, Sage Publications.