Validación de los procesos de determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo

Validation of Processes for Standards Setting for Tests of Educational Achievement

Jesús M. Jornet Meliá
Universitat de València
jornet@uv.es

José González Such
Universitat de València
gonzalej@uv.es

Jesús M. Suárez Rodríguez
Universitat de València
rodrigus@uv.es

Resumen: En este trabajo se presenta una revisión del estado de la cuestión acerca de los métodos para desarrollar procesos de validación de la determinación de Estándares en pruebas de rendimiento educativo. Se analiza el concepto de Validez de los Estándares, así como estrategias de validación, y algunos indicadores de juicio.

Palabras clave: pruebas referidas al criterio, determinación de estándares, puntuaciones de corte; validez de procesos de determinación de estándares.

Abstract: This paper presents a review of the state of affairs concerning methods to develop validation processes of determining performance standards in educational tests. It analyzes the concept of Validity Standards and validation strategies, and some indicators of trial.

Keywords: criterion-referenced tests; determination of standards; cut scores; validity of a standard-setting processes.

Introducción

Los estándares de interpretación de una prueba (en lo sucesivo EE) constituyen la operativización de los criterios de evaluación útiles para valorar, tanto a nivel individual como colectivo, las puntuaciones obtenidas por los sujetos. En un trabajo anterior (Jornet y Suárez, 1989; De la Orden, 2000; Jornet y González Such, 2009) presentamos una breve revisión de los métodos habituales que se han ido desarrollando para determinar los EE, incluyendo procedimientos para la determinación de estándares y puntuaciones de corte. Desde una perspectiva criterial, todos los métodos (aunque por diferentes procedimientos) persiguen "dar cualidad" a las puntuaciones. En definitiva, podríamos decir que los EE están a la base de la emisión del juicio evaluativo que puede emitirse a partir de un instrumento estandarizado y son, por lo tanto, el nexo de unión entre el proceso cuantitativo (medida) y el cualitativo (evaluación, juicio), siendo la expresión de los criterios de interpretación de la evaluación (Ibarra, 2007; Ibarra y Rodríguez-Gómez, 2010).

En la evolución metodológica que se ha dado en este ámbito, se ha ido poniendo de manifiesto la necesidad de establecer procesos que permitieran validar los EE desarrollados para una prueba, tanto si la pretensión es interpretar las puntuaciones individuales como las colectivas –por ejemplo, referidas a evaluaciones de sistemas educativos o instituciones (Tourón, 2009). Ciertamente, las características métricas de calidad de los instrumentos estandarizados (fiabilidad y validez) tienen referentes de diseño y desarrollo diferentes (y ello es así tanto para pruebas de rendimiento como para medidas de otros constructos no necesariamente de carácter cognitivo o independientemente del contexto sociocultural en que se analice –Pérez-Carbonell, Ramos y López-González, 2009; Gargallo, Suárez-Rodríguez y Pérez-Pérez, 2009; Montero, Villalobos y Valverde, 2007–. La fiabilidad toma como punto de partida el modelo de medida que se utilice (sea Teoría Clásica –TC–, de Respuesta al Ítem –TRI– o de la Generalizabilidad –TG–), y se puede asegurar a partir de un trabajo meticuloso dirigido a la selección de ítems, individualmente considerados, y en conjunto como unidad de medida de la prueba; se dirige, por tanto a las características internas del instrumento. Sin embargo, la validez –como ha sido habitual a lo largo de toda la historia psicométrica–, si bien es fundamental como criterio de calidad del instrumento, no puede asegurarse únicamente a partir de los modelos de medida. Esto es así, dado que en definitiva, la validez alude a la calidad de la prueba para representar adecuadamente la realidad que pretende medir, por lo que necesariamente hace referencia a las cualidades de la misma para reflejar de forma fehaciente el fenómeno evaluado; es por este motivo, que la validez se aborda como análisis de la relación entre el instrumento y el fenómeno que

pretende evaluar, y se convierte en consecuencia en un sistema de estudio (o estudios) que permita, desde la investigación, aportar evidencias acerca de ello.

Como señala Hambleton (1984) la validez, en gran medida, depende del uso de las puntuaciones y es en este sentido, por lo que se hace necesario, validar específicamente los procesos que se hayan utilizado para diseñar el sistema de interpretación de las mismas. En consecuencia, diríamos que validar estos procesos constituye una de las evidencias fundamentales acerca de la validez global de las pruebas.

En este trabajo, presentamos algunas consideraciones metodológicas que se han ido proponiendo y utilizando para este propósito.

La validez de la determinación de EE: algunas consideraciones conceptuales

La Validez de la prueba y la determinación de sus EE son dos aspectos íntimamente vinculados, dado que aquélla depende, en gran medida, de la utilización de sus puntuaciones (Jornet, 2008). En esta línea, en los nuevos EE para Tests Psicológicos y Educativos de la AERA, la APA y el NCME (1999) se señala:

> "Un paso crítico en el desarrollo y uso de algunas pruebas es establecer uno o más puntos de corte dividiendo la escala de puntuaciones para dividir la distribución de puntuaciones en categorías […] Los puntos de corte plasman las reglas de acuerdo con las que se usan o interpretan [las puntuaciones] en las propias pruebas. Por lo tanto, en algunas situaciones la validez de las interpretaciones de la prueba podría depender de las puntuaciones de corte" (p. 53).

Y, aunque la consideración siguiente la realizan respecto a las pruebas de certificación, entendemos que es importante tenerla en cuenta a nivel general; así, se indica que "la validez de las inferencias realizadas de la prueba depende de si el estándar para pasar hace una diferencia legítima entre el rendimiento suficiente e insuficiente" (NCME,1999, p.157). Por ello, el modelo que apoye la determinación de los EE debe estar bien seleccionado, diseñado, y llevado a la práctica de manera rigurosa (Kane, 1994, 2001; Ruiz-Primo, Jornet y Backhoff, 2006). Las facetas del concepto de Validez de los EE son las siguientes (Jornet, 2008):

a) las derivadas del concepto de prueba estandarizada,
b) las que están implicadas en el proceso de trabajo que se desarrolla para diseñar el sistema de interpretación de puntuaciones de la prueba, y
c) las que devienen de los usos evaluativos de la prueba.

Las facetas derivadas del concepto de prueba estandarizada, se refieren a la Validez de Constructo y a la de Contenido: dependen, en gran medida, "de la calidad de la definición del dominio educativo (DE) como universo de medida. El estándar debe representar de forma adecuada los elementos relevantes del DE, que resultan clave en la descripción de la calidad del aprendizaje" (Jornet y Perales, 2001, p. 200). Los elementos de juicio son componentes implicados en la calidad del estándar y constituyen la base de toda la prueba. El análisis inicial del contenido a evaluar, el desarrollo de especificaciones y la escritura de ítems, constituyen el elemento básico de calidad de los EE. Por ello, la calidad con que se desarrollen esas fases de la elaboración de las pruebas es clave, tanto para la validez en su conjunto, como para la definición del estándar. Aspectos importantes de estas etapas para la posterior definición del estándar son: a) la especificación precisa de las unidades del universo de medida, b) el acuerdo inter-jueces al respecto, y c) la generalizabilidad[1] de la definición (basada en la diversidad de juicios implicados).

Cuadro 1. Criterios y recomendaciones a considerar en cuanto al proceso de juicio para aportar Credibilidad a los EE

Criterios	Recomendaciones
Para la elaboración de la prueba, lo más adecuado es que se trabaje con *diversos grupos de jueces* –Comités–.	En diversas tareas, como por ejemplo: la definición del DE, elaboración de ítems, revisiones,...[2], de forma que la actuación secuencial de los mismos facilite un trabajo interno que coadyuve a la validez de la prueba, basándose en la diversidad de puntos de vista (generalización) y en el control sucesivo que se ejerce entre comités. En este marco, el comité que propone el Estándar es conveniente que sea el que ha trabajado la especificación del EE en términos de especificaciones y/o ítems.
Composición de los comités	Deben ser representativos de los especialistas del área y nivel al que se dirige la prueba. En este sentido, hay que señalar que deben ser lo suficientemente amplios como para representar a todos los sectores que puedan darse en el ámbito de interés, pero a la vez debe asegurarse la homogeneidad entre los expertos participantes, en cuanto a las características de su actividad (por ejemplo, que sean todos docentes en activo de la materia y nivel), dado que la diversidad de juicio que se da en todos estos procesos debe ser independiente, en todo caso, de factores relativos al origen de los miembros del comité.

[1] El término *generalizabilidad*, en este caso, se usa de forma genérica y no hace referencia a la Teoría de la Generalizabilidad (TG).

[2] Asimismo, es necesario considerar los efectos de interacción social que suelen darse en este tipo de dinámicas (Fitzpatrick, 1989).

Selección de los expertos que componen los comités de desarrollo	Se debe basar en su experiencia y conocimiento de la materia y nivel objeto. Ver Estándar 1.7. de la APA, AERA y NCME (1999).
Adecuación de las técnicas de emisión de juicio	Adecuar el formato de juicio a emitir (Reid, 1991), método de emisión de juicio, pregunta/s de referencia, forma de puntuación, etc... al tipo de Estándar a diseñar. Ver Estándar 4.21. de la APA, AERA y NCME (1999). Es conveniente utilizar rondas sucesivas de emisión de juicio, entre las que se utilicen informaciones de feedback a los jueces. Este tipo de informaciones, pueden estar referidas a las consecuencias de aplicación del estándar (por ejemplo, porcentajes de sujetos en cada categoría o nivel de rendimiento), al grado de acuerdo/discrepancia entre los jueces, a las relaciones de fiabilidad con cada puntuación de corte, etc. En este sentido, "es conveniente abordar el análisis de las consecuencias de la aplicación del estándar propuesto sobre un grupo de estudiantes, que actúe como criterio, de manera que pueda facilitar el feedback de información para el grupo de jueces; ello ayuda a ajustar de forma más realista el estándar (propuesta derivada del método de Jaeger)" (Jornet y Perales, 2001, p. 201).

Respecto a las facetas derivadas de los usos evaluativos, se refieren a las acepciones de validez relativas al impacto y validez del plan de evaluación en que se usan las pruebas en su conjunto: *credibilidad y utilidad*. Ambas acepciones constituyen elementos determinantes para la aceptación de la evaluación y, en consecuencia, si es el caso, disponga de impacto sobre el objeto evaluado; es decir, que *dote a la evaluación de capacidad de cambio* –ver Cuadros 1 y 2–.

La *Credibilidad* se basa en la adecuación del procedimiento de la prueba en general, y del desarrollo del Estándar en particular. En definitiva, depende de la calidad global del proceso técnico seguido para el diseño y desarrollo de todos los componentes de la prueba. Un aspecto de especial relevancia es la adecuación de los comités que hayan intervenido en el desarrollo del estándar. Los criterios y recomendaciones más importantes que afectan a la credibilidad relativas al proceso de juicio se recogen en el Cuadro 1, y las relativas a los procedimientos empíricos o técnicos del proceso, en el Cuadro 2.

Por su parte, *la Utilidad*, hace referencia al grado en que los resultados de la evaluación se usan para la toma de decisiones acerca del sistema educativo. Obviamente, tales decisiones podrían darse a diferentes niveles: a) Orientación de la política educativa, b) Especialistas en diseño curricular, c) Profesorado, y d) Los

evaluados –ver Cuadro 3–. Por otra parte, hay que tener en cuenta que la utilidad es, en definitiva, también una consecuencia de la credibilidad de la evaluación y, a su vez, la retroalimenta. Así, la utilidad necesariamente se apoya en la aceptación sociopolítica de la evaluación y en el uso social de sus resultados.

La comprobación técnica de la utilidad requiere del seguimiento del impacto de la evaluación y en la confirmación de sus resultados. De este modo, puede incluir un amplio abanico de aproximaciones: desde análisis históricos de la realidad educativa, identificando si los cambios políticos y legislativos en el sistema educativo han tenido en cuenta –o se relacionan– con resultados de evaluaciones, pasando por análisis comparados, hasta investigaciones dirigidas a valorar el conocimiento y uso de los resultados y estándares de la evaluación en diferentes estamentos o colectivos (desde los administradores educativos a los docentes); o bien, estudios de carácter técnico en los que se determinen los posibles errores en la evaluación y se hayan tenido en cuenta en el Estándar aquellos que resulten menos lesivos[3]. En cualquier caso, el análisis de la utilidad de los EE debe realizarse a medio y largo plazo, incluyéndolo en aproximaciones más amplias relativas a la utilidad de las evaluaciones.

Las terceras, las que devienen del uso de la prueba, se identifican como evidencias de *Validez Criterial*, *sea Concurrente y/o Predictiva*. Están relacionadas a su vez con la Utilidad de los EE, y su comprobación se sustenta en analizar el valor explicativo del funcionamiento de los EE para diferentes usos evaluativos (Jornet, 2008). Así, la Validación Criterial de los EE, se puede entender como

> "elementos de contextualización del uso de las puntuaciones. Resultan muy importantes respecto a la utilización diagnóstico-evaluativa, en tanto en cuanto tienen valor explicativo del funcionamiento del estándar. En este caso, las estrategias de su consecución son de carácter empírico y deben contener criterios acerca de la valoración de la utilidad de la prueba respecto a las decisiones que debe sustentar" (Jornet y Perales, 2001, p. 200).

[3] Partiendo de que toda evaluación con pruebas estandarizadas contiene necesariamente error, la definición del Estándar y, en consecuencia la puntuación de corte, puede asumirse teniendo en cuenta que siempre se producirá error –aunque lo deseable es que no exista o sea mínimo–. El tipo de error a asumir (la identificación de Falsos Aptos –positivos– o de Falsos No-aptos –negativos–) es elegible al determinar el estándar. Dependerá del uso final de la evaluación, la aceptación de uno u otro tipo de error como menos lesivo.

Cuadro 2. Criterios y recomendaciones de carácter empírico o técnico para aportar Credibilidad a los EE

Criterios	Recomendaciones
El análisis empírico de la prueba así como de la identificación de PC, hay que sustentarlo en modelos sólidos	Preferiblemente TRI de uno, dos o tres parámetros.
Sin embargo, lo anterior debe realizarse respetando los elementos de juicio	Es decir, éste no debe modificarse ni someterse a imperativos de ajuste empírico al Modelo. Así, hay que analizar previamente los requisitos y supuestos de ajuste del Modelo, como por ejemplo la *Unidimensionalidad* de la prueba[4].
Adecuación de las técnicas estadísticas de síntesis del estándar	Basarse en técnicas estadísticas de síntesis de las PC propuestas por los jueces. Es necesario utilizar controles de convergencia de juicios.
En la identificación de las PC, se debe incluir alguna referencia acerca de la fiabilidad en relación con el nivel de habilidad que marca cada una de ellas	Estándar 2.14 de la APA, AERA y NCME (1999): "Cuando los puntos de corte sean especificados para selección o clasificación, los errores estándar de medida deben ser informados respecto a las puntuaciones de corte" (p. 35).
No se debe incluir en los procesos de identificación de las PC, criterios relativos al número de sujetos que deben o pueden superar el estándar. Éste debe depender exclusivamente del nivel de competencia de los sujetos evaluados	Estándar 14.17 de la APA, AERA y NCME (1999): "El nivel del rendimiento requerido para pasar una prueba de acreditación debería depender de los conocimientos y las destrezas necesarias para el rendimiento aceptable en la ocupación o la profesión y no debe ser ajustado con el propósito de regular el número o la proporción de personas que aprobaran la prueba" (p. 162). Este aspecto es extensible a cualquier tipo de prueba, dado que lo que evalúa es la calidad de aprendizaje.

[4] Joaristi y Lizasoáin (2008) y Lizasoáin y Joaristi (2009) demuestran que, si bien todas las pruebas de rendimiento pueden tender hacia la unidimensionalidad, ésta se relaciona estrechamente con la complejidad del curriculum de referencia, por lo que es más factible observar unidimensionalidad en pruebas referidas a niveles educativos inferiores (primaria), que en niveles más elevados (educación secundaria).

Documentación detallada y precisa de todo el proceso	De forma que se facilite su contraste por cualquier agente externo a la entidad u organismo que desarrolle la prueba y los EE. Estándar 4.19 de la APA, AERA y NCME (1999): "Cuando las interpretaciones propuestas consideran uno o más puntos de corte, la lógica y los procedimientos usados para establecer los puntos de corte deben estar claramente documentados" (p. 59).
La cualificación definitiva de niveles debe realizarse contrastando su representatividad respecto de los EE previamente definidos	Se atenderá, pues, a la habilidad o competencia global a que se refiera cada conjunto de ítems asignados como característicos de cada nivel, teniendo en cuenta las descripciones realizadas en la Tabla de Especificaciones, es decir considerando los diversos aspectos que componen el Universo de Medida medidos por la prueba. Si se utiliza algún método que tenga en cuenta las dificultades de los ítems para su asignación a los niveles de rendimiento, es conveniente identificar el grado de dificultad por encima de los cuáles se considera un reactivo característico de un nivel, así como el grado por debajo del cuál debe situarse en el nivel anterior.
Actuación de un comité meta-evaluador	Que revise y aporte una evaluación documentada acerca del conjunto de procesos que afectan al desarrollo de los estándares. Dicho Comité debería estar formado por especialistas de reconocido prestigio, independientes del organismo que realiza la evaluación.
Desarrollo de estudios de validación tanto acerca del proceso de determinación de EE, como de los EE como producto	Siempre que sea posible, se incluirán referencias a los estudios de validación de los estándares (Hambleton, 2001; Kane, 2001). La tipología de estudios a realizar, obviamente, será diferente para cada tipo de estándar. (Ver Estándar 2.15 de la APA, AERA y NCME (1999). *(apartado de Estrategias de validación y criterios de calidad de los estándares).*

En el Cuadro 4 se sintetizan diversas cuestiones relativas a las evidencias que pueden recabarse y que corresponden a las facetas de validez mencionadas en este apartado.

Cuadro 3. Dimensiones de utilidad de los EE

Dimensiones	Consideraciones
La orientación política	En las pruebas que sirvan para evaluar los sistemas educativos, los EE deben permitir orientar la toma de decisiones respecto a aspectos estructurales y funcionales del mismo, tales como el tipo de resultados de sus alumnos, sus escuelas, la evolución del mismo a través del tiempo, o los asociados a cambios en la orientación política del sistema, etc...
El diseño curricular	Se trata de ofrecer información para la mejora del currículum o de su implementación, por lo que deben de representarlo adecuadamente; es decir, no debe haber distancia entre el currículum diseñado y el evaluado, de forma que los EE deben servir para valorar el grado y calidad de implementación del currículum.
La mejora de la actuación docente	Se trata de aportar información precisa que pueda orientar mejoras o innovaciones en la planificación, desarrollo del currículum o en soluciones metodológico-didácticas para el aula.
Los evaluados	En las pruebas dirigidas a evaluar personas, se trata de aportar una evaluación justa y equitativa, que refleje adecuadamente la calidad del aprendizaje que es capaz de realizar cada persona. Implica la independencia de sesgos.

Estrategias de validación y criterios de calidad de los estándares

Teniendo en cuenta los elementos de validez comentados en el apartado anterior, es necesario abordar la *determinación de EE* como un proceso de definición del sistema de interpretación de las puntuaciones de la prueba que, a su vez, debe ser validado.

La *validación de los EE*, "debe entenderse como una acumulación de evidencias de la adecuación del mismo para el propósito de la evaluación" (Jornet y Perales, 2001, p. 200). Para orientar la validación de EE debe atenderse tanto el proceso de determinación de EE, como a los EE como producto (Hambleton, 2001; Kane, 2001; Camilli, Cizek y Lugg, 2001). En el Cuadro 5 se recoge una síntesis de los criterios presentada por Cizek, Bunch y Koons (2004)[5], adaptada de Pitoniak (2003),

[5] Cuyos fundamentos ya se anticipaban en Cizek (2001).

y que constituye una buena panorámica de los diferentes elementos a tener en cuenta en la validación de EE. Los Criterios de procedimiento y los Internos –a excepción de los dos últimos de los internos (Consistencia de la Decisión y Otras medidas)– representan criterios de evaluación del proceso de determinación de EE, mientras que los Externos (y los dos internos antes mencionados) representan aspectos a evaluar acerca de los EE como producto.

Asimismo, presentamos un ejemplo acerca de las estrategias de validación, la síntesis del plan de evaluación para la validación de EE del Modelo de determinación de Niveles de Logro del INEE (Jornet y Backhoff, 2008). Se basa en dos aproximaciones: a) la evaluación del proceso de determinación de EE (ver cuadro 6), y b) la validación de los estándares como producto.

Respecto a la *evaluación de los EE como producto del comité*, hay que señalar que se trata de un área que forma parte de la validación global del instrumento. Cizek, Bunch y Koons (2004) indican que, para valorar los EE como producto, se debe analizar la *racionalidad* y la *replicabilidad* de los EE. La *racionalidad* se puede evaluar por el grado en el que las PC derivadas del proceso de determinación de EE clasifican a los sujetos en grupos de una manera consistente con otra información sobre los mismos. Por ejemplo, se trata de analizar la congruencia entre las clasificaciones de la prueba en comparación con las que aportan otras pruebas o sistemas que puedan asumirse como fiables y válidas[6]. La *replicabilidad* hace referencia a la comparación de la determinación de EE a partir de la utilización de métodos diferentes y/o comités independientes. No obstante, esta última característica tiene problemas difíciles de subsanar, desde los de coste (que son muy elevados) a los derivados de dilucidar, en el caso de discrepancias, los motivos o factores que pueden estar a la base de su explicación, de forma que sería casi imposible decidir acerca de dos métodos bien ejecutados aunque ofrecieran resultados diferentes. En definitiva, se trata de aportar evidencias que avalen los EE definidos como sistema de interpretación de la prueba.

[6] Por ejemplo comparar las clasificaciones de una prueba nacional con las que facilite otra prueba internacional, o bien, analizar si la distribución de sujetos en categorías es consistente con las que se darían en la evaluación habitual de esos sujetos en las clases.

Cuadro 4. Síntesis de evidencias de validez de los EE (Jornet, 2008)

Facetas de Validez	Cuestiones a responder con las evidencias de validación
Faceta 1. Constructo y Contenido	*¿Las descripciones usadas para describir los EE proveen una guía adecuada para interpretar la calidad de ejecución de los estudiantes?* *¿Las descripciones de los niveles establecidos son diferenciales, graduales, y representan diferencias objetivables de calidad de aprendizaje?*
Faceta 2. En relación a la credibilidad del proceso para determinar los EE	*¿El comité de expertos que definen los EE es representativo?* *¿Incluye especialistas en Investigación Educativa, en Curriculum, y profesores?* *¿Es diverso (por tipología de escuelas, estados...)? ¿Los criterios utilizados para seleccionar los expertos han sido pertinentes al problema?* *¿La forma en que los expertos emiten sus juicios ha sido adecuada? ¿La tarea a realizar por los expertos estaba clara? ¿Se ha formado a los expertos en la tarea que debían realizar? ¿Se ha comprobado que habían entendido bien la tarea?* *¿Los expertos han aportado resultados congruentes entre sí? ¿El consenso intersubjetivo ha sido completo acerca de cada nivel de logro? ¿Ha habido dificultades para conseguir dicho consenso? ¿Se ha llegado al consenso de forma natural, sin producirse actuaciones directivas de ninguno de los expertos o de los coordinadores del proceso?* *¿Todos los elementos que se han producido en el proceso de determinación de NL están bien documentados? ¿Hay información de todos los aspectos que se han producido en el proceso? ¿El proceso ha sido transparente? ¿Es posible replicar el proceso?* *¿Hay un un comité meta-evaluador que pueda valorar todos los elementos del proceso y del producto desarrollado?*
Faceta 3. Relacionadas con el uso de la prueba	*¿Se han realizado estudios y recabado evidencias acerca de la adecuación de los EE establecidos? ¿Las descripciones de los EE representan lo que los estudiantes realmente saben y pueden hacer? ¿Son diferenciales en cuanto a la tipología de aprendizaje? ¿Los estudiantes clasificados en un EE alto son exitosos también en la escuela? ¿Los estudiantes clasificados en un EE bajo son menos exitosos también en la escuela?¿Los estudiantes clasificados en un EE alto son exitosos posteriormente?¿Los estudiantes clasificados en un EE bajo son menos exitosos posteriormente?* *¿En qué grado se reflejan adecuadamente las características de los sujetos, las escuelas, los sistemas educativos...?* *¿Las calificaciones derivadas la prueba son equivalentes, a nivel de escuela (no de estudiantes), con las calificaciones escolares aportadas por el profesorado?* *¿En qué grado la prueba refleja el currículum implementado/aplicado? ¿Los EE sirven para orientar innovaciones en la escuela, el curriculum o la política educativa?*

Cuadro 5. Criterios para evaluar los procedimientos de determinación de estándares

(Adaptado por Cizek, Bunch y Koons –2004– de Pitoniak –2003–)

Criterio de evaluación	Descripción
DE PROCEDIMIENTO	
Carácter explícito	Grado en que los propósitos y procesos de la determinación de estándares estaban claros y explícitamente articulados a priori.
Viabilidad	Facilidad de la puesta en práctica de los procedimientos y el análisis de datos; el grado en que los procedimientos son creíbles e interpretables para las audiencias relevantes.
Puesta en práctica	Grado en que los procedimientos eran razonables, y sistemática y rigurosamente desarrollados: selección y formación de los participantes, la definición del nivel de rendimiento y la recogida de información.
Feedback	Grado en que los participantes están seguros del proceso y de la(s) puntuación(es) de corte resultantes.
Documentación	Grado en que las características del estudio se han analizado y documentado para los propósitos evaluativos y comunicativos.
INTERNOS	
Consistencia en el método	Precisión en el cálculo de la(s) puntuación(es) de corte.
Consistencia intrapanelista	Grado en el que un participante es capaz de dar valoraciones o clasificaciones que son consistentes con las dificultades empíricas de los ítems y grado en el que las clasificaciones o valoraciones cambian entre rondas.
Consistencia interpanelista	Consistencia de las valoraciones o clasificaciones de los ítems y puntuaciones de corte entre los participantes.
Consistencia de la decisión	Grado en que la aplicación repetida de la(s) puntuación(es) de corte produciría clasificaciones consistentes de los sujetos examinados.
Otras medidas	Consistencia de las puntuaciones de corte a través de los tipos de ítems, áreas de contenido y procesos cognitivos.
EXTERNOS	
Comparaciones con otros métodos de determinación de estándares	Consistencia de las puntuaciones de corte a través de las réplicas usando otros métodos de determinación de estándares.
Comparaciones con otras fuentes de información	La relación entre las decisiones tomadas usando la prueba y otros criterios relevantes (por ejemplo, notas, rendimiento en pruebas que miden constructos similares, etc.).
Razonamiento de las puntuaciones de corte	Medida en que las recomendaciones de las puntuaciones de corte son viables y objetivas (incluyendo valoraciones pasa/no pasa y el impacto diferencial sobre subgrupos relevantes).

La acumulación de evidencias acerca de la calidad de los EE puede requerir un tiempo que normalmente no se tiene, dado que es necesario disponer de un sistema de interpretación de los EE para poder utilizar y comunicar los resultados. Cualquier estudio que implique seguimiento, o que dilate en exceso el poder ofrecer resultados, puede constituir un problema de difícil solución para la utilización evaluativa de las pruebas. *¿Qué garantías podríamos considerar suficientes para comenzar a utilizar los EE?* Hay que asumir que en el mismo proceso de determinación de los EE se han de obtener esas garantías. Para ello, es preciso que los indicadores que se hayan utilizado para identificar las PC sean pertinentes al problema. La dificultad radica en que el trabajo con comités de estas características es un área emergente y existen pocas alternativas de análisis para valorar la concordancia de juicio y seleccionar las PC (Jornet, 2008). Para ello se pueden utilizar diversos indicadores de convergencia de los juicios (Castro, 2001). Por ejemplo, la Precisión de Juicio (PJ), la Razón de Acuerdo (RA) o los Sesgos de las Valoraciones (SV), para valorar la identificación de cada PC; así como aplicaciones de indicadores de convergencia global, como la W de Kendall; o las valoraciones de los jueces respecto a la seguridad y calidad de las PC identificadas. –ver Cuadro 7–.

Cuadro 6. Síntesis del plan de evaluación del proceso de determinación de Niveles de Logro de los EXCALE del INEE-México (Jornet y Backhoff, 2008)

¿Qué se evalúa?	Se pretende aportar información respecto a dos aspectos: a) Si el proceso de elaboración de EE es adecuado, y b) Si los EE propuestos como resultado de este proceso tienen garantías de calidad suficientes como expresión del consenso intersubjetivo de los participantes en el mismo.
¿Para qué se evalúa?	La *finalidad de la evaluación* es poder analizar si el proceso de determinación de los EE se ha llevado a cabo de forma adecuada. El uso de la evaluación, en este caso, es doble: a) *formativo*, de manera que durante el proceso se trata de recabar información para corregir los problemas detectados durante el desarrollo del mismo que puedan afectar a la calidad y validez de los niveles de desempeño identificados, y b) *sumativo*, como evidencia de validez, acerca de la representatividad y calidad de los niveles de desempeño identificados como sistema de interpretación de puntuaciones de la prueba.
¿Qué audiencias están implicadas?	Pueden distinguirse dos grandes grupos de audiencias: a) Los *participantes en el proceso de determinación de EE*, que son los que los desarrollan y aportan, a su vez, información para la evaluación (miembros de comités de desarrollo de los EE, coordinadores del proceso y evaluadores externos), y b) Los *receptores finales de la información* (profesionales de la educación –administradores, profesores...–, decisores políticos y sociedad en general.

¿Quién realiza la evaluación?	La lleva a cabo un Equipo Evaluador externo, y un Comité meta-evaluador, cuya función es analizar si la evaluación –y sus conclusiones– está bien realizada.
¿Qué variables e indicadores se tienen en cuenta?	De Entrada: Características profesionales de los participantes. De Proceso: Comprensión de la tarea y de los procedimientos a utilizar, Número de sesiones de juicio, Cambios en la identificación de puntuaciones de corte de una a otra sesión de juicio. Cambios en la fiabilidad asociada a las puntuaciones de corte de una a otra sesión de juicio, Cambios en la distribución porcentual de los sujetos a partir de los niveles de logro identificados de una a otra sesión de juicio. De Resultado: Satisfacción con el proceso de formación, Satisfacción con los procedimientos utilizados, Satisfacción con el funcionamiento global del comité, Congruencia en la identificación de puntuaciones de corte (en cada sesión de juicio), Perspectivas univariada y multivariada, Fiabilidad (función de información) asociada a las puntuaciones de corte en cada nivel, Distribución porcentual de los sujetos en los EE, Satisfacción con la adecuación de los EE determinados. De Contexto: Comparación del funcionamiento de los diversos comités de las diferentes materias. Continuidad lógica de los EE en una materia para diversos niveles educativos (Escalación vertical –si fuera posible– entre niveles educativos diferentes). Análisis lógico de los niveles de logro identificados para cada materia con los utilizados en otro proyecto evaluativo que resulten comparables.
¿Quién aporta la información?	En la evaluación se pretende triangular la información que proviene de diversas fuentes de información. Así, se recoge información de: Los *participantes*, en al menos tres aspectos: a) el análisis de sus respuestas de identificación de puntuaciones de corte, b) el conocimiento y comprensión de los métodos y procedimientos que se utilizan, y c) sus opiniones acerca del proceso. El *coordinador de prueba*: sus valoraciones acerca del proceso. El *observador externo*: sus valoraciones acerca del proceso. El *Comité meta-evaluador* (CME): sus valoraciones metodológicas (validación del informe de evaluación).
¿Qué instrumentos se utilizan?	*Cuestionarios* –para recabar opiniones de los participantes en el proceso–, *Hojas de registro de observaciones* –para recoger las observaciones que realizan los coordinadores de la prueba que actúan como coordinadores de este proceso–, y *hojas de registro de PC* –en donde se consigne la PC seleccionada por cada participante del Comité 2, así como sus valoraciones acerca de los obstáculos ocurridos para su identificación–.

Cuadro 7. Síntesis de indicadores para la evaluación de la convergencia en las PC

Indicadores	Formas de estimación
Precisión del juicio (PJ)	En cada ronda de juicio se valora la desviación de los juicios respecto a la Mediana como indicador base para valorar la distancia de los juicios emitidos a la PC seleccionada. Este indicador tiene como referente para orientar el criterio el valor mismo de la desviación de la escala (100 puntos, en caso de usar TRI), de forma que puede entenderse que una $\sigma = 0$ indica convergencia total de juicios.
Razón de acuerdo (RA)	Entre jueces al determinar una puntuación de corte. Se estima como el porcentaje de jueces que coinciden en la identificación de una puntuación. Pueden tomarse diferentes intervalos para valorar la coincidencia de juicios[7]: $\sigma \pm 2.5\%$, $\sigma \pm 5\%$, $\sigma \pm 7.5\%$, $\sigma \pm 10\%$, $\sigma \pm 12.5\%$, $\sigma \pm 15\%$,..., $\sigma \pm 25\%$. En cada caso se contabilizan los jueces que emiten valoraciones en cada rango. Teniendo en cuenta los resultados que se evidencian en la literatura especializada, se considera un buen nivel de convergencia una $RA \geq 10\%$, al menos en un intervalo de $\sigma \pm 10\%$.
Sesgos de valoraciones (SV)	Se estiman las distancias medias que se producen por encima y por debajo de la puntuación de corte a fin de analizar las tendencias de valoración que se han producido al determinar la puntuación de corte, con el fin de valorar su robustez final. Entre las PC no convergentes se entiende que una identificación es insesgada cuando las distancias entre juicios superiores e inferiores son equidistantes, y se asume que está sesgada cuando se desvía por efecto de la presencia de valoraciones extremas por encima o por debajo de la PC.
Concordancia global de la serie de PC identificadas	En casos de EE politómicos se puede analizar la convergencia general entre los jueces en la serie completa de PC identificadas.
Valoraciones de los participantes	Acerca de la representatividad de los niveles obtenidos en cuanto al porcentaje de sujetos identificados en cada nivel, así como sus opiniones sobre el proceso de identificación de cada PC. Estas informaciones se sintetizan mediante análisis estadísticos descriptivos.

Con todo, estas aproximaciones se refieren al acuerdo entre los jueces en cada PC en su conjunto. *¿Qué relación existe entre los ítems y los EE? ¿A cada nivel le corresponde*

[7] La σ se refiere a la variabilidad de los juicios realizados acerca de la identificación de las PC.
[8] Aunque puede resultar problemático en cuanto a la métrica.

realmente un conjunto ítems característicos? Este es otro problema que se puede abordar en el momento de hacer disponible la prueba para su uso. Desde una perspectiva multivariada se pueden identificar algunas aproximaciones que coadyuven como evidencias de validez de los EE. Desde una *perspectiva exploratoria*, podría abordarse el análisis de perfiles de desempeño, a partir de un análisis de conglomerados de k-medias[8]. Se trata de explorar perfiles de ejecución en la prueba y analizar si dichos perfiles son congruentes con los EE establecidos. Un requisito de calidad de estos perfiles es que sean escalados, es decir, que cada perfil presente un nivel de ejecución diferencial en todos los ítems respecto de los demás perfiles, y que presente diferencias estadísticamente significativas en todos los ítems. Por otra parte, y desde una *perspectiva explicativa*, y considerando que cada ítem presenta un nivel diferencial de ejecución para cada EE, puede analizarse mediante regresión logística si los ítems característicos de cada nivel son los que se definen al establecer los EE. En este sentido, la calidad de la explicación constituirá el criterio de bondad necesario para asumir los EE como sistema de interpretación (Jornet, 2008).

No obstante, la validación no puede darse por concluida una vez esté acabada de diseñar la prueba. Se deberá continuar recogiendo evidencias de validez durante el uso evaluativo de las pruebas. Pueden realizarse diversos estudios que avalen o refuten los EE diseñados; estudios que pongan de manifiesto la capacidad predictiva de los EE, basados en aproximaciones de carácter longitudinal. Implicaría realizar estudios de correlación y regresión múltiple en los que se pudiera poner de manifiesto el valor de los EE para predecir la calidad del aprendizaje en diferentes colectivos.

A MODO DE CONCLUSIÓN

Los procesos de validación de determinación de EE de interpretación de las puntuaciones de las pruebas se está convirtiendo en un área emergente de trabajo. Es fundamental para asegurar la asegurar la calidad de la interpretación de puntuaciones, y requiere de un trabajo minucioso que podría situarse en el ámbito de la investigación evaluativa, sustentado en una perspectiva de complementariedad metodológica –cuantitativa/cualitativa–, al necesitar procedimientos de trabajo de evaluación de las situaciones en que se dan este tipo de procesos. Situaciones que requieren del juicio de expertos, observación, etc. El camino está apuntado, pero es necesario dedicar una mayor atención al desarrollo de métodos que podamos utilizar con esta finalidad.

Recepción del original: 3 de marzo de 2009
Recepción de la versión definitiva: 16 de marzo de 2010

Referencias

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

Camilli, G., Cizek, G.J. y Lugg, C.A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts*, *Methods*, *and Perspectives* (pp. 445-476). Mahwah: Erlbaum.

Castro, M (2001). How accurate are writing performance assignment raters? *2001 LAUSD rater reliabity study. CSE Technical Report. California:* CRESST: UCLA.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31-50.

Cizek, G.J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G.J. Cizek (Ed.), *Setting performance standards: Concepts*, *Methods*, *and Perspectives* (pp. 3-17). Mahwah: Erlbaum.

De la Orden, A. (2000). Estándares en la evaluación educativa. Ponencia presentada en las primeras *Jornadas de Medición y Evaluación*. Valencia: Universidad de Valencia.

Fitzpatrick, A.R. (1989). Social influences in standard setting: The effects of social interaction on group judgments. *Review of Educational Research*, *59*, 315-328.

Gargallo, B., Suárez-Rodríguez, J.M. & Pérez-Pérez, C. (2009). El cuestionario CEVEAPEU. Un instrumento para la evaluación de las estrategias de aprendizaje de los estudiantes universitarios. *RELIEVE. Revista Electrónica de Evaluación Educativa*, 15, 2. Extraído el 24 de enero de 2010 de http://www.uv.es/RELIEVE/v15n2/RELIEVEv15n2_5.htm

Hambleton, R.K. (1984). Validating the test scores. En R.A. Berk (Ed.), *A guide to criterion-referenced test construction.* Baltimore: Johns Hopkins University Press.

Hambleton, R.K. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L.N. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I* (pp. 97-114). Washington: Council of Chief State school Officers.

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards: Concepts*, *Methods*, *and Perspectives* (pp. 89-116). Mahwah: Erlbaum.

Hambleton, R.K., Jaeger, R.M., Plake, B.S. y Mills, C.N. (2000). *Handbook for setting standards on performance assessment.* Washington: Council of Chief State School Officers.

Ibarra, M. S. (2007) (Dir.): *Proyecto SISTEVAL. Recursos para el establecimiento de un sistema de evaluación del aprendizaje universitario basado en criterios, normas y procedimientos públicos y coherentes.* Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.

Ibarra, M.S. y Rodríguez-Gómez, G. (2010). Aproximación al discurso dominante sobre la evaluación del aprendizaje en la universidad. *Revista de Educación*, *351*, 385-407.

Joaristi, L. y Lizasoáin, L. (2008). Estudio de la dimensionalidad empleando análisis factorial clásico y análisis factorial de información total: análisis de pruebas de matemáticas de primaria (5º y 6º cursos) y secundaria obligatoria. *RELIEVE. Revista Electrónica de Evaluación Educativa*, *14*(2). Extraído el 13 de enero de 2010 de http://www.uv.es/RELIEVE/v14n2/RELIEVEv14n2_2.htm

Jornet, J.M. y González Such, J. (2009). Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre Educación*, *16*, 102-123.

Jornet, J.M. y Backhoff, E. (2008). *Modelo para la determinación de Niveles de Logro y Puntos de Corte de los Exámenes de la Calidad y el Logro educativos (Excale).* México: Instituto Nacional de Evaluación Educativa (INEE). Extraído el 15 de octubre de 2010 de http://www.inee.edu.mx/images/stories/Publicaciones/Cuadernos_investigacion/30/Completo/cuaderno30a.pdf

Jornet, J.M. (2008). La validación de los procesos de determinación de NL en las pruebas de desempeño. Ponencia presentada en el *VIII Foro de Evaluación Educativa.* Mérida (México): Instituto Nacional de Evaluación Educativa (INEE) / Centro Nacional para la Evuluación de la Educación Superior (CENEVAL).

Jornet, J.M. y Perales, M.J. (2001). La interpretación de puntuaciones en las pruebas de rendimiento: elementos metodológicos en el desarrollo de estándares. En CENEVAL, *Quinto Foro de Evaluación Educativa* (pp. 195-204). Ensenada: Centro Nacional para la Evaluación de la Educación Superior (CENEVAL).

Jornet, J.M. y Suárez, J.M. (1989). Revisión de modelos y métodos en la determinación de estándares y en el establecimiento del punto de corte en evaluación referida a criterio (ERC). *Bordón*, *41*(2) 277-301.

Kane, M.T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Kane, M.T. (2001). So much remains the same: Conception and status of validation in setting standards, In G.J. Cizek (Ed.), *Standard performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Lewis, D.M., Mitzel, H.C., Green, D.R. y Patz, R.J. (1999). *The bookmark standard setting procedure.* Monterey, CA: McGraw-Hill.

Lizasoáin, L. y Joaristi, L. (2009). Análisis de la dimensionalidad en modelos de valor añadido: estudio de las pruebas de matemáticas empleando métodos no paramétricos basados en TRI. *Revista de Educación*, *348*, 175-194.

Montero, E, Villalobos, J. y Valverde, A. (2007). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la Universidad de Costa Rica: un análisis multinivel. *RELIEVE*, 13, 2. Extraído el 14 de enero de 2010 de http://www.uv.es/RELIEVE/v13n2/RELIEVEv13n2_5.htm.

Pérez-Carbonell, A., Ramos, G. y López-González, E. (2009). Diseño y análisis de una escala para la valoración de la variable clima social aula en alumnos de educación primaria y secundaria. *Revista de Educación*, 350, 221-252.

Pitoniak, M.J. (2003). *Standard setting methods for complex licensure examinations.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice*, *10*(2), 11-14.

Ruiz-Primo, A., Jornet, J.M. y Backhoff, E. (2006). *Acerca de la Validez de los exámenes de la calidad y el logro educativos (Excale).* México: Instituto Nacional de Evaluación Educativa (INEE). Extraído el 15 de octubre de 2010 de http://www.inee.edu.mx/index.php/component/content/article/3666

Tourón, J. (2009). El establecimiento de estándares de rendimiento en los sistemas educativos. *Estudios sobre Educación*, *16*, 127-146.