Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo

En este trabajo se presenta una revisión del estado de la cuestión acerca de los métodos para determinar estándares en pruebas de referencia criterial. Se analiza el concepto de estándar, sus componentes y los problemas implicados en las tareas de juicio en estos métodos. Se revisan los procedimientos para diseñar descriptores de estándares y los métodos para identificar puntuaciones de corte.

Palabras clave: pruebas referidas al criterio, determinación de estándares, puntuaciones de corte, rendimiento educativo.

EE016

Jesús M. Jornet Meliá

Departamento de Métodos de Investigación y Diagnóstico en Educación. Facultad de Filosofía y Ciencias de la Educación. Universitat de València
jornet@uv.es

José González Such

Departamento de Métodos de Investigación y Diagnóstico en Educación. Facultad de Filosofía y Ciencias de la Educación. Universitat de València
gonzalej@uv.es

Criterion-referenced Evaluation: Standards Settings of Interpretation for Educational Achievement Tests

IIn this paper, a review of the state of the art about the methods for Standard setting in Criterion-Referenced Tests is presented. We analyze the concept of Standard, its components, and the problems implied in the tasks of judgment in these methods. The procedures to design descriptors for standards and the methods for identifying cutting scores are reviewed.

Keywords: criterion-referenced tests, standard-setting, passing scores, educational achievement.

1. Introducción

La Evaluación Criterial (EC) se ha venido desarrollando desde los años sesenta del siglo pasado hasta la actualidad. Su advenimiento

se produce a partir de las consideraciones de Robert Glaser, entre las que destaca la inadecuación de los sistemas normativos de construcción de tests psicométricos para la elaboración de pruebas estandarizadas de rendimiento. El factor principal que sustenta estas posiciones es la necesidad de que las pruebas de rendimiento educativo se puedan interpretar en función de criterios absolutos de calidad del aprendizaje, dado que las decisiones que se toman a partir de ellas así lo requieren.

Los temas fundamentales de desarrollo de las pruebas criteriales son dos: a) el análisis y especificación del Dominio Educativo (DE) como universo de medida desde el que se originan las pruebas, y b) el desarrollo de estándares (EE) o sistema de interpretación de puntuaciones dirigido a establecer un juicio de valor acerca de la calidad del aprendizaje. En conjunto, en los procesos de construcción de pruebas criteriales, se han ido imponiendo procesos que diferencian claramente a éstas respecto de las tradicionales. En especial, hay que resaltar como punto central el hecho del trabajo colegiado de profesionales –del área que mida la prueba– como referente esencial para asegurar la calidad de la misma y de sus componentes. Así, el énfasis se ha ido poniendo sobre la participación de expertos tanto en la definición del DE –el diseño, revisión y selección de ítems–, como en la definición de los EE. Una característica adicional a resaltar es que todo el proceso de elaboración de la prueba debe estar bien integrado, dado que si no se dispone de un adecuado análisis y especificación del DE, difícilmente se pueden llegar a interpretar de forma adecuada los puntajes de las pruebas.

Su impacto ha sido muy importante en la Medición y Evaluación Educativas en general, si bien, sus usos de mayor trascendencia se han identificado en el ámbito de las pruebas estandarizadas, en especial, las de certificación y/o admisión, y en las pruebas a gran escala dirigidas a la evaluación de instituciones y sistemas educativos. En este último caso, hay que señalar que buena parte de los sistemas de evaluación nacionales e internacionales han ido adoptando los principios de elaboración de pruebas criteriales, tanto en lo que se refieren a pruebas alineadas al currículum, a estándares de calidad de referencia, a competencias, etc.

En este trabajo nos centramos sobre el segundo tema de interés señalado: el sistema de interpretación de puntuaciones, y específicamente sobre los métodos que se han ido proponiendo para este cometido.

2. Componentes de los EE de interpretación de puntuaciones

Como señala De la Orden (2000): "En la evaluación de resultados educativos los estándares se identifican con el modelo de aprendizaje de los alumnos y determinan el conjunto de decisiones de selección, clasificación, calificación y promoción a todos los niveles del sistema" (p. 2).

En el concepto de EE[1], en el sentido en que aquí lo utilizamos, hay que diferenciar entre *estándares* y *puntuaciones de corte*. A este respecto, algunos autores (Van der Linden, 1980; Jornet, 1987; Jornet y Suárez, 1989; Kane, 1994; Cizek, 2001; Cizek, Bunch y Koons, 2004) hemos abogado por diferenciar ambos términos. El término estándar se reservaría para hacer referencia al sistema de criterios de interpretación, la definición teórica de los niveles de desempeño, logro o competencia, mientras que el término puntuación de corte (PC) indicaría la puntuación en la prueba que sirve para diferenciar entre dos niveles de desempeño. En cualquier caso, EE y PC son dos componentes de un mismo proceso. En la definición de los EE podemos identificar diversos componentes –ver Cuadro 1–. En el Cuadro 2 se muestran ejemplos de etiquetas usuales en EE.

Cuadro 1.

Componentes de definición de los EE (Jornet y Backhoff, 2008)

Componente	Descripción
Categorías o etiquetas relativas a niveles de desempeño	Generalmente politómicos (con 3 a 6 categorías), se describen por etiquetas alusivas al nivel de dominio o simplemente con números.
Descriptores de los niveles de desempeño	Relativas al tipo de aprendizaje característico de los sujetos clasificables en cada nivel.
Puntuaciones de corte	Las puntuaciones que en la prueba sirven para diferenciar entre cada uno de los niveles de desempeño.
Ítems característicos	Ítems que son capaces de realizar los sujetos de un determinado nivel de desempeño.

Según plantea Linn (1994) –ver Cuadro 3– se pueden diferenciar diversos tipos de EE. Dependiendo del propósito a que se dirijan los mismos, del plan de evaluación en que se utilice (el tipo de decisiones a apoyar), y la orientación que sigan los participantes en el proceso, los EE pueden diferir considerablemente. Así, el trabajo previo a su determinación debe estar enfocado a aclarar de forma muy precisa la tipología de EE que necesita la institución que desarrolla la evaluación.

[1] Acepción alternativa del término EE, sería la que se da en Evaluación Educativa relativa a las Normas de calidad con las que juzgar los planes de evaluación de programas (por ejemplo, los estándares establecidos por el *Joint Committee on Standards for Educational Evaluation*, 1994) –inicialmente publicados en 1981–, o en Medición Educativa, la referida a las descripciones de calidad de aprendizaje usuales para el desarrollo de pruebas referidas a estándares. En nuestro caso, EE lo asumimos como sinónimo de niveles de desempeño, ejecución, logro o competencia.

Cuadro 2.

Ejemplos de etiquetas de niveles de rendimiento

Origen	Etiquetas
National Assessment of Educational Progress	Básico, Competente, Avanzado.
Terranova (2ª ed.) (CTB / McGraw Hill)	En camino, Progresando, Cerca de la competencia, Competente, Avanzado.
Pruebas de rendimiento del Estado de Ohio	Limitado, Básico, Competente, Acelerado, Avanzado.
Estado de California, Tests de California	Muy por debajo del nivel básico, Debajo del nivel básico, Básico, Competente, Avanzado.
Estado de Texas, estándares de valoración de Texas de conocimientos y destrezas	No llega al nivel usual, Llega al nivel usual, Rendimiento destacado.
INCE (Instituto Nacional de Calidad y Evaluación)[2]: Diagnóstico del sistema educativo español, 1998	Sin etiquetas, identificación mediante la cualificación de la escala numérica.
Proyecto PISA[3]	Niveles numéricos.
SERCE (LLECE)	Cuatro niveles numéricos.
EXCALE (INEE, México)	Por debajo del nivel básico, Básico, Medio, Avanzado.

(Adaptado desde Cizeck, Bunch y Koons, 2004, p. 34).

[2] Actualmente IE (Instituto de Evaluación).

[3] En el Informe 2000, utiliza cinco niveles numéricos para Lectura, y tres (Máximo, Medio, Mínimo) para Matemáticas y Ciencias. En 2003, en todas presenta cinco niveles numéricos. En 2006, Matemáticas y Ciencias cuentan con seis niveles.

3. Consideraciones acerca de los métodos para la determinación de EE y PC

Revisar de forma detallada los métodos de determinación de EE es una tarea amplia y prolija, que excede los límites razonables de esta presentación. Sin embargo, es necesario realizar algunas consideraciones respecto a las características de los métodos que se han ido desarrollando para este propósito. Para este cometido, analizaremos: a) la evolución de los métodos de determinación de EE, b) unas reflexiones acerca de la arbitrariedad y consenso intersubjetivo, c) consideraciones respecto a las aproximaciones para definir las categorías de contenido de los EE, d) tipos de métodos para identificar PC, y e) criterios para seleccionar el método de determinación de estándares.

Cuadro 3.
Tipología de estándares (Linn, 1994)

Tipo de estándar	Descripción
Exhortación	Representan metas deseables de logro a las que debe tender la mejora de un sistema educativo o de los estudiantes.
Ejemplificación o muestra de rendimiento	Representan las habilidades o competencias características de diversos niveles de ejecución.
Rendición de cuentas para educadores	Representan metas curriculares precisas, orientando la evaluación hacia el contraste entre el currículum diseñado, el implementado y los logros educativos.
Certificación del logro del estudiante	Identifican un nivel mínimo de competencia del logro de estudiantes.

3.1. Evolución de los métodos de determinación de EE

En este apartado, nos centraremos en descripciones genéricas de las características y problemas de los métodos de determinación de EE, así como en las valoraciones y argumentos que han ido guiando su evolución. En la evolución de estos métodos podemos identificar tres grandes momentos (Jornet y Backhoff, 2008):

a) Los procedimientos de determinación de EE se plantean y desarrollan en el ámbito de los *tests criteriales* (TRC), en el siglo XX, desde la década de los 60[4]. La problemática que

[4] En 1963 Robert Glaser publica su artículo *Instructional technology and the measurement of learning outcomes: some questions*, en el que se plantean las bases de desarrollo de los tests criteriales.

se plantea es poder desarrollar métodos que permitieran aportar una valoración absoluta de calidad de las realizaciones que hacían los estudiantes en las pruebas[5], dado que la interpretación de puntuaciones basada en las normas de grupo constituyen una falacia de base para el tipo de juicios que deben tomarse en Educación, como señalamos en la Introducción de este artículo. De hecho, este tipo de interpretaciones es similar a las que hace cualquier profesor acerca de las realizaciones de sus estudiantes, si bien, se trata de encontrar métodos que permitan identificar criterios de forma objetiva.

b) Hasta la década de los 80, los métodos desarrollados se dirigen a interpretaciones de tipo dicotómico (pasa/no-pasa), vinculados a usos de pruebas de admisión y/o certificación, aunque también se identifican propuestas para usos más restringidos para pruebas de uso en el aula.

c) En los años 90 se comienzan a aportar métodos dirigidos a determinar EE politómicos, utilizados en el marco de pruebas a gran escala dirigidas a la evaluación de sistemas educativos. El *National Assessment of Educational Progress* (NAEP) fue uno de los primeros en expresarlos a partir de series graduadas de niveles de desempeño: *Basic* –básico– *Proficient* –competente– y *Advanced* –avanzado– (Cizek et al., 2004). En España, el estudio sobre el Diagnóstico del Sistema Educativo Español de 1998 (De la Orden et al., 1998) identifica niveles de competencia a partir de los ítems característicos de cada uno de ellos, a partir de su comportamiento empírico. Finalmente, los estudios internacionales –como por ejemplo Proyectos PISA o SERCE– también han adoptado sistemas politómicos para informar de sus resultados.

En la actualidad, resulta un sistema frecuente de información de resultados. Los problemas metodológicos han evolucionado también, desde los referidos a la orientación general de este tipo de procesos (por ejemplo, el rol de las tareas de juicio frente a las empíricas) hasta problemas más específicos (por ejemplo, las técnicas de trabajo con jueces o los indicadores de convergencia de juicio) que ponen de manifiesto una mayor madurez de este ámbito metodológico).

3.2. Arbitrariedad y consenso intersubjetivo

Un problema que ha estado presente, y aún lo está, en el desarrollo de este ámbito metodológico ha sido el rol del juicio de expertos en la determinación de los niveles de desempeño. Por su interés, nos ha parecido importante dedicarle estas líneas.

En esta problemática aún subyace la ya clásica polémica planteada a partir del importante trabajo de Glass (1978) en el que señalaba la arbitrariedad de los procedimientos que se

[5] El estudiante domina o no el contenido educativo, logra o no los objetivos, es competente o no lo es.

habían ido desarrollando, basados en juicio[6]. Pese a las limitaciones que él indicaba, se ha ido imponiendo el hecho de la necesidad de los procesos de juicio para determinar los niveles de desempeño. Lo cierto es que el concepto de calidad es diverso –entre personas y a través del tiempo– y que, en cualquier caso, tiene un componente claro: su carácter subjetivo. La subjetividad es inherente al concepto de calidad (De la Orden, 2000; Jornet, 2008). De este modo, para afirmar que los niveles de aprendizaje que presentan los estudiantes son de una determinada calidad, no podemos basarnos únicamente en la descripción de las distribuciones empíricas de las pruebas (*planteamiento normativo*), sino que se requiere establecer un juicio de valor acerca de lo que demuestran los sujetos en las pruebas (*planteamiento criterial*).

No obstante, este problema no es exclusivo de las pruebas de rendimiento como instrumento de medida. Por ejemplo, pensemos en un termómetro. Lo que aporta el termómetro es una graduación de la temperatura, pero si deseáramos diseñar un sistema de interpretación que permitiera aportar un juicio acerca de si "hace frío o no", indefectiblemente deberíamos recurrir a las opiniones de las personas. Es obvio que la variabilidad de la percepción de la temperatura es importante entre personas; así, las que viven habitualmente en ambientes muy cálidos perciben antes el frío que otras que provienen de ambientes más fríos. Sin embargo, ante ello no sería esperable que se plantearan problemas acerca de la "calidad de la interpretación", y se asumiría que ésta depende de las percepciones personales. Sin duda, sería así porque este tipo de juicios tiene menos trascendencia que los que se pretenden tomar a partir de las pruebas, y el instrumento de medida es más adecuado para escalar la temperatura que una prueba para escalar el desempeño educativo.

Con todo, si se necesitara interpretar una escala de temperatura, reduciendo su información en términos de la percepción de frío/calor, deberíamos atender al juicio de personas para identificar a partir de qué temperatura se puede afirmar que se da una u otra situación. Es decir, deberíamos identificar la PC, como grados a partir de los cuáles las personas perciben una u otra situación (hace frío, hace calor). Para que este juicio fuera representativo, habiendo controlado de forma precisa la graduación de la temperatura y sus cambios sucesivos, no bastaría con que le preguntáramos a una o varias personas sin control, sino que deberíamos atender de forma precisa la composición del grupo de personas a consultar, asegurando su representatividad (zonas geográficas de origen, edades, sexo, situación física en el momento de la experiencia y demás variables) y, en todo caso, el sistema de EE quedaría limitado para su

[6] En respuesta a las posiciones de Glass, se publicaron diversos trabajos. Entre ellos, destacan los de autores como Popham, Block, Hambleton, Shepard, o Berk. Actualmente la aceptación de los procesos de juicio está asumida y generalizada entre los especialistas en este tema.

uso en el contexto en que se hubiera desarrollado. El criterio para identificar en cuántos grados se percibe el cambio de frío a calor (o viceversa), debería establecerse por procedimientos que recogieran el consenso intersubjetivo del grupo. Así, la calidad de la interpretación se basaría en (para asegurar la validez y fiabilidad): el control exhaustivo de la experiencia, la calidad del grupo consultado (lo que apoyaría sus posibilidades de generalizar los resultados), y los procedimientos que hayamos seguido para identificar el consenso intersubjetivo –juicio– (como criterio de calidad que nos alejara de la arbitrariedad de la interpretación).

Este mismo problema y sus líneas de solución son los que han orientado el desarrollo de los métodos de determinación de EE en las pruebas de rendimiento educativo. El problema, en todo caso, será llegar a métodos que permitan establecer interpretaciones de calidad de aprendizaje que no sean arbitrarias, sino que estén fuertemente ancladas en la realidad. En este sentido, las posiciones que se han ido imponiendo resaltan la necesidad, utilidad y valor de las tareas de juicio en estos procedimientos, e incluyen las cautelas propias que aconsejan tener las limitaciones que se han ido poniendo de manifiesto. El consenso intersubjetivo, y el modo en que se ha llegado a éste, constituyen la garantía de calidad de la interpretación.

3.3. Acerca de las aproximaciones para definir las categorías de contenido de los EE

Las categorías de contenido, las descripciones de lo que son capaces de realizar los sujetos de cada nivel de desempeño y la selección de ítems característicos para cada uno de ellos, son componentes de los EE que se desarrollan mediante procesos de juicio –ver Cuadro 4–. Por ello, la definición de calidad de aprendizaje es el componente más cualitativo del proceso de determinación de EE. Los enfoques que se han dado para abordar esta problemática se pueden clasificar en: a) dependientes exclusivamente del análisis del DE, b) dependientes del DE y del funcionamiento de la prueba (mixtos), y c) dependientes fundamentalmente del comportamiento empírico en la prueba.

En el primer caso (enfoque a), actuarían como referentes desde los que se pueden desarrollar los ítems y, por tanto, se pueden diseñar al comienzo del desarrollo de la prueba. Las ventajas de este procedimiento son: a) permiten abordar el diseño de pruebas cuando no existe un currículum único, o bien, cuando éste es muy difuso[7]; y b) permiten un buen muestreo de ítems para representar de forma adecuada tareas representativas de cada nivel de desempeño, de forma que se pueden diseñar ítems que discriminen adecuadamente entre niveles. Por el contrario, como desventaja más importante, se puede citar que se corre el riesgo de plantear EE descontextualizados de la realidad, por lo que se requiere una comprobación posterior de tipo empírico.

[7] En el diseño de pruebas para la evaluación de competencias, sería la orientación más pertinente.

Una variación de este tipo de acercamientos es aquélla que se basa en el DE, pero en la que se realizan los EE en un proceso mixto juicio-empírico, de forma que su diseño se sustenta sobre el análisis del DE, pero se informa adicionalmente del comportamiento empírico de los ítems. Este doble trabajo, si bien ofrece mayores posibilidades de realismo a los EE finales, puede poner de manifiesto las lagunas de la prueba acerca de la falta de ítems que permitan discriminar en algunos de los niveles establecidos. Por ello, es recomendable realizarlo cuando se dispone de datos del pilotaje de la prueba, para poder corregir los problemas detectados, aunque en muchas ocasiones se realizan ya con los datos finales de la misma.

Por último, los que dependen sólo del comportamiento empírico de la prueba, se realizan habitualmente con los datos de la prueba definitiva. El procedimiento se basa en analizar los ítems característicos de cada nivel de habilidad total en la prueba y establecer cortes en la escala, de forma que se diseña el descriptor de cada nivel a partir de los ítems que componen cada nivel.

Cuadro 4.

Ejemplo de descriptor de un nivel de competencia (NAEP para pruebas de lectura de 4º Grado)

	Descripción del nivel de rendimiento ***avanzado***
Descriptor genérico	*Los estudiantes de cuarto grado que están en el nivel avanzado deben poder generalizar sobre los tópicos en la selección de lecturas y demostrar un conocimiento suficiente acerca de cómo los escritores componen y usan las estrategias literarias. Cuando leen textos apropiados para cuarto grado, deben poder juzgarlos de forma crítica y, en general, dar respuestas minuciosas que demuestren que han comprendido el texto.*
Ejemplos basados en ítems característicos	Por ejemplo, cuando leen textos literarios los estudiantes de nivel avanzado deben poder hacer las generalizaciones sobre lo relevante de la historia y prolongar su significado integrando las experiencias personales y las otras interpretaciones con las ideas indicadas por el texto. Deben poder identificar los recursos literarios como la lengua figurada. Cuando leen textos informativos los alumnos de cuarto grado de nivel avanzado deben poder explicar el propósito del escritor usando material de soporte del texto. Deben poder hacer juicios críticos sobre la forma y el contenido del texto y explicar sus juicios claramente.

3.4. Tipos de métodos para identificar puntuaciones de corte (PC)

Han sido múltiples los intentos realizados hasta la fecha para tratar de exponer y valorar los métodos para abordar este problema (Ziecky, 1995, 2001). Se han presentado diversos sistemas de clasificación de los mismos (Meskauskas, 1976; Glass, l978; Shepard, 1980, 1984;

Berk, 1986; Jornet, 1987; Jornet y Suárez, 1989; Cizeck, 1996a; Hambleton, Jaeger, Plake, y Mills, 2000; Cizeck et al., 2004). Para este trabajo, seguimos una tipología simple que ya hemos presentado en un trabajo anterior (Jornet y Backhoff, 2008):

- *Métodos de juicio*, donde revisamos los métodos basados en el juicio que realizan expertos acerca de los ítems, los sujetos o las tareas.
- *Métodos empíricos*, donde reseñamos metodologías que se basan prioritariamente en el comportamiento empírico de la prueba.
- *Métodos mixtos*, donde revisamos los métodos que conjugan el juicio de expertos con otras informaciones de carácter empírico.

3.4.a. Métodos de juicio

En este apartado, integramos la revisión de tres grandes conjuntos de métodos, que se diferencian en función del objeto sobre el que se realiza el juicio: a) sobre los ítems, b) sobre los sujetos, y c) sobre las tareas.

Entre los denominados *Métodos de Juicio*, destacan los métodos basados en el *juicio sobre los ítems*, como los de Nedelsky (1954), Angoff (1971), Jaeger (1978) o Ebel (1962, 1972). Estos métodos se basan en la idea de *sujeto limítrofe*, que se define como aquél que obtiene una puntuación media (o mediana) entre los dos grupos que pueden considerarse como criterio: aptos/no-aptos. Para identificar esa puntuación límite (o punto de corte –*passing score*–) se parte del análisis lógico de los ítems que componen la prueba. La tarea que, en general, deben realizar los jueces en la aplicación de estos métodos es muy compleja, dado que se trata de evaluar cuál sería el comportamiento esperado de un sujeto limítrofe entre dos niveles de competencia ante cada ítem de la prueba. De esta forma, no se trata de una mera estimación del rendimiento. Por este motivo una de las precauciones básicas que hay que tomar en su aplicación (Livingston y Zieky, 1982) es contrastar la estructuración efectuada por los jueces con los datos obtenidos a partir de una aplicación empírica, de forma que si no son convergentes es necesaria su revisión.

Pese a que la mayor parte de métodos, tal como fueron propuestos originalmente, no se utilizan en la actualidad (sino que han sido objeto de variaciones y actualizaciones), bien es cierto que algunas de sus modificaciones se han instaurado como las soluciones más viables para este propósito[8]. Los problemas que se han ido argumentando al respecto, así como sus líneas de solución actual, se recogen en el Cuadro 5.

[8] Como por ejemplo, el de Angoff o el de Jaeger.

Cuadro 5.

Problemas y soluciones desarrolladas en torno a los métodos de juicio sobre los ítems

Problemática	Comentarios y soluciones
La complejidad cognitiva que se planteaba a los jueces, desde la idea de sujeto limítrofe, hasta el formato de juicio (valorar alternativas de los reactivos, etc.).	La mayor parte de propuestas actuales enfatizan la simplicidad en el formato de juicio como un elemento clave para el éxito en la aplicación de este tipo de métodos.
La limitación de las propuestas a ítems de alternativas, de forma que no contemplaban aplicaciones para ítems de desarrollo.	Las variaciones actuales suelen ajustar el formato de juicio[9] respecto al *ítem* considerándolo globalmente, de forma que se abren las posibilidades de aplicación a cualquier tipo de *ítem*.
El hecho de que fueran métodos diseñados para identificar estándares dicotómicos (pasa/no pasa).	Se planteaban como métodos para aplicaciones muy limitadas, de forma que solo podrían dar respuesta para aquellas pruebas en las que se requiriera un juicio simple, como la admisión a un programa o el egreso del mismo. Actualmente, los esfuerzos se vierten hacia la determinación de EE politómicos.
La variabilidad entre las PC propuestas por cada juez utilizando un mismo método.	Se han podido ir superando a partir de estrategias de análisis de datos más refinadas, de forma que pueden sintetizarse las PC a partir de métodos robustos, y aplicaciones de técnicas para detectar jueces que ofrecen valoraciones extremas.
La enorme variabilidad entre los EE producidos por métodos alternativos.	Aunque en algunos estudios comparativos se han identificado las fuentes de variabilidad, lo cierto es que aún en la actualidad no se ha identificado una explicación que aclare esta problemática.
La elevada exigencia que habitualmente se observaba en los EE producidos por este tipo de procedimientos.	Ello se relaciona con dos grandes factores: la formación de los expertos que participan en los comités y el nivel de exigencia usual en los profesores al valorar lo que se debe aprender. La superación de este problema se basa en formar previamente a los jueces y en la propuesta de Jaeger (1978) de introducir retroalimentación de información acerca de las consecuencias de la aplicación de los EE producidos. Se basa en aportar información –entre sesiones de juicio consecutivas– acerca de cuáles serían los efectos de aplicar el estándar, así como en señalar los elementos de discrepancia, etc.

[9] Por ejemplo, el formato de juicio podría ser como el que sigue: Para superar la prueba, *¿sería necesario que un sujeto respondiera adecuadamente este ítem?*

Qué estrategias utilizar para formar a los participantes en los comités, así como qué tipo de información ofrecer como retroalimentación a los participantes en los comités, y de qué manera hacerlo, se han convertido en centros de interés para los investigadores de esta área (Raymond y Reid, 2001; Reckase, 2001), que han impactado en otros métodos de juicio o mixtos.

Por otra parte, otro grupo de métodos que ha tenido también buena acogida y trascendencia ha sido el de *métodos basados en el juicio sobre sujetos*. Sistematizados inicialmente por Livingston y Zieky (1982), han tenido una amplia aplicación y uso. Dentro de esta categoría se podrían incluir diversos tipos de métodos, siempre y cuando tomen como referencia una valoración externa a la prueba acerca de la capacitación de los sujetos (tomados individualmente o en grupos), tales como el de Validación de Grupos Criterio (Berk, 1976, 1980) –ver tabla 7–, o el método basado en el Grupo de Referencia (Livingston y Zieky, 1982)[10]. Así, se podría decir que estos métodos son en definitiva un procedimiento de identificación (y a veces de validación) de las PC basado en una evaluación pormenorizada de los sujetos que se asume como criterio. Se requiere dos tipos de datos sobre cada sujeto: 1) la puntuación en la prueba, y 2) el juicio sobre la adecuación del conocimiento y habilidades del sujeto en relación al DE.

La dificultad básica de este tipo de procedimientos radica en la fiabilidad y validez de la selección de sujetos que deben actuar como criterio. Tanto la valoración individual de los sujetos por procedimientos alternativos a la prueba, como la formación de un juicio global acerca de un grupo (como por ejemplo, instruidos/no instruidos) tiene problemas. En el primer caso, a la falta de seguridad en la identificación de los sujetos (los que en realidad están capacitados y los que no lo están), se suma el elevado coste del procedimiento. Téngase en cuenta que la calidad del método radica, en todo caso, en la calidad del criterio. En el segundo caso, el procedimiento es más operativo, pues el juicio se realiza acerca de un grupo. Así, por ejemplo, si se desea determinar una PC para el final de un programa, se puede tomar como referencia el nivel de los sujetos que ya lo han superado y ponerlo en relación con los que aún no lo han hecho. Sin embargo, ello tampoco es un elemento de seguridad acerca de la capacitación real de los sujetos. Un problema añadido es decidir acerca del criterio y del procedimiento estadístico que sintetice esa relación capacitados/no capacitados. Por ello, son métodos que pueden tener más utilidad y aplicabilidad en pruebas dirigidas al aula que en pruebas a gran escala. No obstante, esta opción creemos que sigue siendo atractiva para inspirar estudios de validación de EE, más que como procedimiento para la identificación de las PC.

Una evolución metodológica que podríamos situar entre los dos conjuntos de métodos descritos –los de juicio sobre ítems y los de juicio sobre sujetos– son los de *juicio sobre tareas*,

[10] Como los métodos de grupos contrastados, o el método del zig-zag (Up and Down).

también denominados *métodos holistas* (Cizeck et al., 2004). Este tipo de métodos se dirigen a valorar de manera global la tarea de cada sujeto, de forma que a partir de las evaluaciones que realiza un comité de expertos se puedan extraer las PC. Son especialmente útiles en casos de tareas de desarrollo –como por ejemplo, composiciones escritas, tareas artísticas y similares– o en casos en que en la misma prueba hay una gran variedad de tipologías de ítems –y/o tareas–. Como en los casos anteriores, en esta categoría también se incluyen diversos métodos, como el de juicio analítico de Plake y Hambleton (2001), el método de selección de trabajos de Loomis y Bourque (2001), el método "The body of work method" (cuerpo del método de trabajo) propuesto por Kingston, Kahl, Sweeney y Bay (2001).

En términos generales, estos métodos proceden a partir de un comité que revisa y valora una muestra de trabajos de los sujetos examinados. Esta valoración persigue clasificar los trabajos en categorías de rendimiento, bien en las categorías propias de los niveles, bien en categorías que representen los límites entre niveles. Las ventajas de este tipo de métodos es que realizan un análisis bastante preciso del procedimiento de juicio a utilizar, el modo en que se identifican las puntuaciones de corte, la utilización de procesos de retroalimentación informativa para los participantes en los comités, etc.; así como son los métodos más apropiados para la identificación de niveles de tareas de desarrollo. No obstante, son métodos costosos dado que requieren que un panel grande de expertos valore una muestra importante de trabajos.

3.4.b. Métodos empíricos

El siguiente gran grupo de *métodos* es el que podríamos denominar *empíricos*. En este caso, agrupamos métodos de diversa índole que tienen en común el hecho de que la mayor parte del procedimiento se sustenta sobre información empírica. No se trata de procedimientos exentos de tareas o elementos de juicio, sino que en la mayor parte de los casos, el peso de la información empírica respecto a los elementos de juicio es mucho mayor. Se caracterizan por: 1) todos ellos utilizan la escala de puntuaciones observadas y en la misma expresan el punto de corte resultante, y 2) son procedimientos empíricos en el sentido de que tienen en cuenta la información distribucional empírica que se da en la muestra estudiada.

Identificamos tres grandes grupos de métodos: a) los modelos de estado, b) los modelos continuos basados en la teoría de la decisión y c) los basados en la distribución de los ítems sobre la escala de habilidad total.

En el primer conjunto de procedimientos, se pueden identificar propuestas que no han llegado a tener trascendencia práctica o que ésta ha sido muy limitada, como es el caso de

[10] Como los métodos de grupos contrastados, o el método del zig-zag (Up and Down).

los modelos de estado de Roudabush, el de Emrick y Adams y Emrick, presentados todos ellos en la década de los 70 y revisados por Macready y Dayton (1980). Estos modelos arrancan de una concepción del Aprendizaje "todo/nada", es decir, se posee o no la habilidad o dominio en cuestión. Este punto de partida es consistente con el énfasis que se tuvo en la Evaluación Referida al Criterio respecto a la definición de unidades de dominio discretas homogéneamente definidas (Shepard, 1984).

Otro grupo de métodos de interés fueron los *Modelos continuos basados en la Teoría de la Decisión*. En contraposición a los modelos de estado ya comentados, estos modelos suponen la existencia de una variable latente continua sobre la cual se debe determinar el punto de corte de tal modo que se optimicen los resultados de la decisión. Pero, en sentido estricto, este conjunto de modelos no deben entenderse como procedimientos para la determinación de un estándar. Así, Van der Linden (1980) explicitó claramente que:

> "[…] la aproximación, basada en la teoría de la decisión, a los TRCs no es una técnica para el establecimiento de estándares sino una técnica para minimizar las consecuencias de los errores de medida y de muestra, los cuales, preferentemente formando parte de una rutina normal, se deberán seguir cada vez que se use una técnica de establecimiento de estándares" (p. 470).

Es decir, una vez determinado un estándar, punto de corte en puntuaciones verdaderas, se podrá aplicar un procedimiento basado en la teoría de la decisión para determinar una puntuación que minimice los efectos derivados de los errores, punto de corte en puntuaciones observadas[11]. En este sentido, "las técnicas basadas en la Teoría de la Decisión no son sustitutas de los métodos de establecimiento de estándares sino que se deben de utilizar cada vez que se ha usado uno de estos métodos para tomar decisiones basadas en datos de tests que contienen error" (Van der Linden, 1984, p. 11). Pese a ello, en su momento dieron origen a diversos procedimientos, que se diferenciaban básicamente en la consideración del error (Jornet, 1987; Jornet y Suárez, 1989).

Entre los métodos basados en la vinculación de ítems con la escala de habilidad total, se pueden encontrar diversas aplicaciones de carácter empírico. Se trata de identificar los ítems característicos de cada nivel de habilidad total, basándose en un nivel alto de probabilidad de respuesta al ítem (por ejemplo, igual o superior al 67%), sobre una escalación con Teoría de Respuesta al Ítem (TRI). Asignados los ítems, se analizan los puntos de inflexión en la escala, de forma que se establece para un rango determinado de habilidad cuáles son los ítems que responden adecuadamente los sujetos de ese rango de manera diferencial respecto

[11] Van Der Linden (1980, 1984) insiste en denominar estándar al punto de corte expresado en puntuaciones verdaderas y punto de corte al punto de superación expresado en puntuaciones observadas. Esto, según el autor, clarifica notablemente el propósito de cada método expuesto en el presente punto).

a los clasificables en los rangos adyacentes. Una referencia de utilidad para este propósito es el uso del Mapa de Wright. La cualificación de los EE se realiza a posteriori, es decir, a la luz de los contenidos o competencias a que se refieren los ítems asignados a cada rango (Backhoff, Peón, Andrade y Rivera, 2006). Son procedimientos especialmente útiles para aquellos casos en que lo que se pretende medir con la prueba no tiene un referente curricular preciso y, en todo caso, conceptualmente se asemeja a un constructo teórico no observable similar a las variables psicológicas, en el sentido de las variables de producto educativo mediato descritas por De la Orden (1985). Por ejemplo, variables como la Expresión Oral o Escrita, evaluadas a partir de rúbricas de calificación.

3.4.c. Métodos mixtos

Un grupo de métodos que tuvo gran atractivo en su momento fue el denominado como *Métodos de Compromiso*. En ellos se pretende establecer la PC a partir de un acuerdo entre los niveles mínimos de competencia estimados por jueces y la distribución empírica resultante de la ejecución del grupo de referencia. Entre ellos, se pueden identificar los métodos de De Gruijter (1985), Hofstee (1983) y el de Beuck (1984). Descritos por Shepard (1984) y Cizek (1996a, 1996b), la concepción de los tres métodos es similar y su objetivo el mismo. Básicamente difiere en el modo en que se establece el acuerdo entre ambas fuentes de información (lógica y empírica). Probablemente la mayor ventaja que presentan estos métodos es que parten de una base de sentido común, dado que respetan los niveles mínimos que objetivamente deben considerarse en el terreno educativo, junto al hecho de que tienen en cuenta la distribución empírica de los resultados de la prueba (Jornet y Suárez, 1989). Su diseño original es para pruebas de admisión, por lo que la razón de pase es un criterio más que no tiene por qué identificarse en otro tipo de pruebas. Los elementos de crítica a estos métodos han sido: la escasa atención al procedimiento de juicio y la justificación estadística de la elección del procedimiento para establecer el compromiso. Ninguno de los métodos parece disponer de justificaciones mejores que los otros en este sentido, según reconocen los mismos autores o los revisores mencionados.

Otra forma de intentar conciliar la estimación lógica con la distribución empírica se identifica en los *métodos de correspondencia de ítems*. Uno de los métodos entre los que actualmente tienen mayor impacto es el *método Bookmark o del marcador*. Presentado por Lewis, Mitzel y Green (1996) y Lewis, Mitzel, Green y Patz (1999), se ha utilizado ampliamente en educación K-12. El método parte de un cuadernillo de ítems ordenados por su dificultad empírica. La tarea de los jueces es identificar el/los ítems que actúan como punto de inflexión entre dos niveles de desempeño previamente definidos por juicio. Lo cierto es que la tarea cognitiva que se plantea a los jueces es mucho más simple que los habituales procedimientos de juicio, así como permite ajustes más realistas de los EE al tener como referencia la dificultad empírica de los ítems. Adicionalmente, en este método –así

como en la mayor parte de propuestas actuales– se cuidan todos los detalles del proceso de emisión de juicios, formatos, probabilidades a considerar en el juicio, forma en que se establecerá la PC, etc. de manera que es un método que si bien no es ideal, puede considerarse que aporta soluciones viables –y sobre todo realistas– para la mayor parte de los problemas señalados.

Una variación o especificación de este método es el Modelo de determinación de niveles de logro de los EXCALE del Instituto Nacional para la Evaluación de la Educación (INEE) de México (Jornet y Backhoff, 2008). Se basa en el trabajo que desarrollan dos comités de forma sucesiva. El primero, compuesto por especialistas en currículum e investigación educativa, los cuáles diseñan los descriptores de cada categoría o nivel de logro a partir de la especificación del universo de medida realizado para el desarrollo de la prueba[12]. El segundo comité, compuesto por profesores en ejercicio, identifica las PC en la prueba que separan los niveles de logro, en un proceso iterativo de emisión de juicio y retroalimentación acerca de las consecuencias de aplicación de las PC identificadas. Finalmente, el primer comité revisa y ajusta el descriptor en función de las PC, y concluye el diseño del descriptor integrando referencias de ejemplo a ítems característicos de cada nivel de desempeño. Todo el proceso se desarrolla basándose en protocolos de actuación muy precisos que guían el trabajo de todos los miembros implicados en el mismo. Así, los protocolos del método son: protocolo para la formación de los comités, para el trabajo del comité 1 –elicitación de descriptores–, del Comité 2 –emisión de juicios, identificación de PC, y retroalimentación de información– y para la validación del proceso y del producto de los comités.

3.5. Criterios para seleccionar el método de determinación de EE

Un problema adicional, dada la enorme oferta metodológica existente es: *¿Qué método elegir?* Como señalan los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1999), "no hay un único método para determinar puntos de corte para todas las pruebas o para todos los propósitos" (p. 53). Junto a este problema hay una realidad que tranquiliza: la evolución de los métodos, así como los estudios comparativos realizados al respecto, al menos ofrecen criterios claros que pueden ayudar a centrar el método a elegir. A este respecto, revisamos las etapas para la determinación de EE, descritas por Hambleton (1998, 2001) –ver Cuadro 5–, como una buena síntesis de los mismos, así como remitimos al lector al apartado relativo a la validez de los EE.

Aunque cada elemento de los señalados por Hambleton es de sumo interés, nos gustaría destacar como componentes clave: la selección y composición del comité de expertos, su

[12] Se parte de un análisis reticular del diseño del curriculum así como de las tablas de especificaciones de los ítems, y se apoyan en el cuaderno de reactivos ordenados.

formación, la elección del método de emisión de juicios y la validación del proceso (a través de un plan de evaluación), como garantías necesarias para que el proceso sea adecuado.

Cuadro 6.
Etapas para el desarrollo de EE (Hambleton, 1998, 2001)

1. Seleccionar un comité de expertos grande y representativo, como base de la validez y fiabilidad de los EE.
2. Elegir el método de determinación de estándares; preparar materiales de formación y el programa de reuniones para la determinación de EE.
3. Preparar las descripciones de las categorías de rendimiento.
4. Formar a los participantes en el uso del método de determinación de EE.
5. Recopilar clasificaciones de ítems y otras valoraciones de los participantes y producir información descriptiva/resumen u otra realimentación para los participantes.
6. Facilitar la discusión entre participantes de la información descriptiva/resumen inicial.
7. Realizar una segunda sesión de clasificaciones/valoraciones; compilar la información y facilitar la discusión como en los pasos 5 y 6.
8. Dar una oportunidad final a los participantes de examinar la información y llegar a los EE finales de rendimiento recomendados.
9. Llevar a cabo una evaluación del proceso de determinación de EE, recogiendo información sobre la confianza de los participantes en el proceso y los EE de rendimiento resultantes.
10. Reunir la documentación del proceso de determinación de EE y cualquier otra evidencia de la validez de los estándares de rendimiento resultantes.

4. A modo de conclusión

La determinación de EE de interpretación de las puntuaciones de las pruebas constituye un área de trabajo ineludible si se desea utilizar las pruebas estandarizadas de rendimiento como indicadores de calidad del aprendizaje. Aunque los problemas implicados en esta tarea son difíciles de solucionar de forma satisfactoria, lo cierto es que los esfuerzos que se han ido realizando en el desarrollo de métodos han sido grandes y aportan en la actualidad soluciones razonablemente aceptables. En la base de todos ellos, se pone de manifiesto la necesidad de identificar el consenso intersubjetivo como referencia precisa para el diseño de EE y como garantía de calidad de los mismos. Este hecho sitúa el diseño y desarrollo de EE en un ámbito de complementariedad metodológica (cuantitativa/cualitativa) y pone de

manifiesto nuevas áreas de interés para el desarrollo metodológico: modelos para conducir el proceso con comités de juicio, técnicas para el análisis de juicios, métodos de validación y evaluación de EE –aunque estos aspectos deben ser objeto de atención en otro trabajo–.■

Fecha de recepción del original: 3 de noviembre de 2008
Fecha de recepción de la versión definitiva: 14 de enero de 2009

Referencias

- American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing. Washington*, DC: American Educational Research Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. En R. L. Torndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Eduction.
- Backhoff, E., Peón, M., Andrade, E. y Rivera, S. (2006). *El aprendizaje de la expresión escrita en la Educación Básica en México. Sexto de primaria y tercero de secundaria*. México D.F.: INEE.
- Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 4*, 4-9.
- Berk, R. A. (1980). *A guide to criterion referenced tests construction*. Baltimore: The Johns Hopkins University Press.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Education Research, 56*(1), 137-172.
- Beuck, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*, 147-152.
- Cizek, G. J. (1996a). Standard setting guidelines. Educational Measurement: *Issues and Practice, 15*(1), 13-21.
- Cizek, G. J. (1996b). Setting passing scores. Educational Measurement: *Issues and Practice, 15*(2), 20-31.
- Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Erlbaum.
- Cizek, G. J., Bunch, M. B. y Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.
- De Gruijter, D. N. (1985). Compromise methods for establishing examination standards. *Journal of Educational Measurement*, *22*, 263-269.
- De la Orden, A. (1985). Hacia una conceptualización del producto educativo. *Revista de Investigación Educativa, 3*(6), 271-284.
- De la Orden, A., Bisquerra, R., Gaviria, J. L., Gil, G., Jornet J. M., López Freire, F. A. et al. (1998). Los resultados escolares. *Diagnóstico del sistema educativo, 1997*. Madrid: Ministerio de Educación y Cultura, Secretaría General de Educación y Formación Profesional, INCE.
- De la Orden, A. (2000, Marzo). Estándares en la evaluación educativa. Ponencia presentada en las primeras Jornadas de Medición y Evaluación, Universidad de Valencia, Valencia.
- Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement, 22*, 15-25.
- Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Glaser, R. (1963). Instructional technology and the measurement of learning out-comes: Some questions. *American Psychologist, 18*, 519-521.
- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237-261.
- Hambleton, R. K. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. En L. N. Hansche (Ed.), *Handbook for the development of performance*

standards: Meeting the requirements of Title I (pp. 97-114). Washington, DC: Council of chief state school officers.

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

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

- Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. En S. B. Anderson y J. S. Helmick (Eds.), *On educational testing* (pp. 109-127). San Francisco, CA: Jossey-Bass.

- Jaeger, R. M. (1978). *A proposal for setting a standard on The North Caroline High School*. Paper presented at the spring meeting of the North Caroline Association for Research in Education, Chapell Hill.

- Joint Committee on Standards for Educational Evaluation (1994). *Standards for evaluations of educational programs, projects, and materials*. New York: MacGraw-Hill.

- Jornet, J. M. (1987). Una aproximación teórico-empírica a los métodos de medición de referencia criterial. Tesis doctoral no publicada, Universitat de Valencia, Valencia.

- 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, Yucatán (Mérida), México.

- 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). *Colección Cuadernos de Investigación*, 30. México D.F.: INEE.

- 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.

- Kingston, N. M., Kahl, S. R., Sweeney, K. y Bay, L. (2001). Setting performance standards using the body of work method. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

- Lewis, D. M., Mitzel, H. C. y Green, D. R. (1996, Junio). Standard setting: A bookmark approach. En D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

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

- Linn, R. L. (1994). *The likely impact of performance standards as a function of uses: From rhetoric to sanctions*. En Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments (pp. 267-276). Washington, DC.

- Livingston, S. A. y Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

- Loomis, S. C. y Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175-218). Mahwah, NJ: Erlbaum.

■ Macready, G. B. y Dayton, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement, 4*, 493-516.

■ Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 46*(1), 133-158.

■ Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 1*4(1), 3-19.

■ Plake, B. S. y Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

■ Raymond, M. R. y Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum.

■ Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task. The role of feedback regarding consistency, accuracy, and impact. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159-174). Mahwah, NJ: Erlbaum.

■ Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*, 447-467.

■ Shepard, L. A. (1984). Setting performance standards. En R. A. Berk. (Ed.), *A guide to criterion-referenced test construction* (pp. 169-198.). Baltimore: Johns Hopkins University Press.

■ Van der Linden, W. J. (1980). Some thoughts on the use of decisión theory to set cutoff scores: Comment on de Gruijter and Hambleton. *Applied Psychological Measurement, 8*, 9-17.

■ Van der Linden, W. J. (1984). Decision models for the use with criterion-referenced tests. *Applied Psychological Measurement, 4*, 469-492.

■ Ziecky, M. J. (1995). A historical perspective on setting standards. En *Proceedings of joint conference on standard setting for large-scale assessments* (pp. 1-38). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

■ Ziecky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19-52). Mahwah, NJ: Erlbaum.

