Una reseña sobre la validez de constructo de pruebas referidas a criterio

Yolanda Edith Leyva Barajas*

El presente artículo es una reseña de los conceptos y métodos que han guiado la construcción de pruebas educativas a gran escala para una variedad de propósitos en la actualidad. Inicia con antecedentes de la evaluación referida a criterio y se desarrolla el tema de la metodología para la construcción de pruebas criteriales, destacando la importancia de la definición del dominio a evaluar y del establecimiento de estándares o puntos de corte que permitan diferenciar niveles de competencia dentro de este dominio, para lo cual se revisan los métodos desarrollados y las tendencias actuales. Se concluye con una reflexión sobre la importancia de establecer líneas de investigación que garanticen la confiabilidad y validez de estas pruebas, para asegurar que las inferencias que se hacen a partir de sus resultados, sirvan como indicadores de calidad del aprendizaje y como apoyo a las decisiones de mejora para las que fueron construidas.

Palabras clave

Evaluación criterial
Competencias
Estándares
Puntos de corte
Validez
Pruebas educativas

This article is a report on the concepts and methods that have guided the construction of large-scale educational tests for several purposes. It starts with precedents for evaluation criteria and the methodology to construct criteria tests. The article highlights the relevance of the definition of the domain to be evaluated, and the establishment of standards that allow differential levels of competence in this domain. To do so, the methods developed are reviewed as well as the current tendencies. It concludes with a reflection about the importance of establishing research lines that guaranty the reliability and validity of these tests, so it can be ensured that the inferences that are made from the results are useful as quality learning indicators and as a support for improving decisions, which are the reason these tests were designed.

Keywords

Evaluation criteria
Competencies
Standards
Cut points
Validity
Educational tests

Recepción: 6 de abril de 2010 | Aceptación: 9 de junio de 2010

* Doctora en Educación. Licenciada en Psicología y maestra en Análisis Experimental de la Conducta por la Facultad de Psicología de la UNAM. Actualmente es directora de evaluación y certificación en el Instituto Internacional de Investigación de Tecnología Educativa, S.C.; miembro de la Red Iberoamericana de Investigadores sobre la Evaluación de la Docencia. Catedrática invitada de la Universidad de Valencia. Últimas publicaciones: (2009), Jesús M. Jornet y Yolanda E. Leyva (coords.), *Conceptos, metodología y profesionalización en la evaluación educativa*, México, INITE; (2010), "La evaluación como recurso estratégico para la mejora de la práctica docente ante los retos de una educación basada en competencias", *Revista Iberoamericana de Evaluación Educativa*, vol. 3, núm. 1, pp. 232-245. CE: yolanda.leyva@inite.edu.mx

Introducción

La importancia de este artículo radica en brindar un marco conceptual y metodológico para ayudar a comprender qué son las pruebas criteriales, cómo se construyen, cómo y para qué se aplican, cómo se fundamenta la interpretación de los resultados que se obtienen a partir de ellas y cuáles son sus principales usos; es decir, para qué sirven. Este tipo de pruebas, tanto internacionales como nacionales, se han venido aplicando en México en las últimas décadas y aún se sabe poco de ellas inclusive dentro del ámbito educativo. Entre las pruebas criteriales más representativas que se aplican en nuestro país, están los Exámenes Generales para el Egreso de Licenciatura (EGEL) del Centro Nacional para la Evaluación de la Educación Superior (CENEVAL); los exámenes de la Calidad y el Logro Educativos (EXCALE) del Instituto Nacional para la Evaluación de la Educación (INEE); y la prueba para la Evaluación Nacional de Logro Académico en Centros Escolares (ENLACE) que aplica la Secretaría de Educación Pública. Entre las internacionales está el Programa para la evaluación internacional de alumnos (*Programme for International Student Assessment*, PISA por sus siglas en inglés), y la prueba del Segundo Estudio Regional Comparativo y Explicativo (SERCE) del proyecto coordinado por la oficina regional de educación de la UNESCO para América Latina y el Caribe, entre otras.

El desarrollo del campo de las pruebas criteriales en México es relativamente reciente y existen sólo algunas experiencias documentadas acerca de estudios de validez apropiados a este tipo de pruebas. En la literatura de países con una gran tradición en su uso para evaluación a gran escala, como es el caso de Estados Unidos, en los primeros años de aplicaciones (de los sesenta a los ochenta), todavía algunos evaluadores asumían que la validez de las mediciones de una prueba criterial consistía sólo en demostrar de manera formal la validez de contenido, concediendo menor importancia a las condiciones bajo las cuales se aplicaban, al uso que se hacía de sus resultados y a las decisiones que dependían de estos resultados. El panorama actual ha evolucionado hacia la creación de nuevas líneas de investigación educativa orientadas a la validación de una gran variedad de pruebas criteriales vinculadas a otras de las ciencias cognoscitivas.

Una revisión de la publicación acerca de los estándares de calidad para la construcción de pruebas psicológicas y educativas publicado en 1999 por la *American Psychological Association* (APA), la *American Educational Research Association* (AERA) y el *National Council on Measurement in Education* (NCME) permite dimensionar la importancia de incluir programas de investigación en la agenda de la evaluación en nuestro país, ya que en estos documentos se destaca que la consideración fundamental de cualquier proceso de evaluación es la validez. En esta publicación se atienden aspectos de la validez de pruebas con referencia a criterio considerando los diversos ángulos del problema de sesgo que pueden afectar los derechos de las personas y la equidad. Se atienden también otros avances técnicos recientes de especial interés para los organismos que aplican pruebas a gran escala, como es el caso de integrar el uso de teorías de medición más actualizadas, por el tipo de información que proporcionan para la mejora continua de estos instrumentos.

De acuerdo con estos estándares de calidad, la validez depende de factores tales como la intención específica de la prueba, el procedimiento usado en su construcción, las condiciones de la colección de los datos, y los procedimientos de enjuiciamiento y medición; así como del análisis de los procedimientos empleados y las características de las personas que proporcionan los datos y la información. Por todo ello se pueden identificar fuentes de evidencia que permiten aclarar diversos factores que pueden afectar la validez (AERA, APA, NCME, 1999).

Como parte de esta reflexión sobre la validez, conviene considerar el reto que representa la evaluación de competencias, lo cual demanda la medición de un constructo teórico más general, lo que a su vez implica la incorporación de una red teórica más amplia y comprensiva. Esto conduce a abordar el problema de la validez no sólo en términos de una o más correlaciones con criterios, como en la práctica tradicional de validez predictiva y concurrente, o en términos de juicios del grado en el cual se ha logrado representar un dominio, como en la práctica común de validez de contenido. El campo de la validez se integra, como ya lo habían referido Cronbach (1971) y Messick (1975), con prueba de hipótesis y con todos los medios empíricos y filosóficos mediante los cuales se evalúan las teorías científicas.

Se ofrece también una reflexión acerca de las implicaciones que tiene esta orientación del concepto de validez para la construcción, administración y uso de pruebas criteriales, ya que los investigadores se han orientado a tender un vínculo entre la validación de pruebas educativas y el campo de investigación de la psicología cognoscitiva, como es el caso de las líneas de investigación propuestas por Shavelson y Ruiz-Primo (2000) para obtener validez cognoscitiva, las cuales prometen evidencias más sólidas para la generalización de los resultados en el ámbito de la evaluación de competencias. Un ejemplo de este tipo de investigación es presentado por Leyva (2004) al integrar un marco teórico acerca de estudios de investigación que se han realizado para validez de constructo de la evaluación de competencias médicas y que ejemplifican el tipo de vínculos con líneas de investigación de las ciencias cognoscitivas, además de una propuesta de estudios de investigación realizados a partir de los resultados de la aplicación del EGEL de Medicina del CENEVAL, para la validación tanto de la definición y estructura del dominio de la prueba, como de los puntos de corte para diferenciar niveles de desempeño o competencia.

Antecedentes de la evaluación referida a criterio

Existe coincidencia en la literatura de evaluación a gran escala, en situar a Glaser como el autor del primer artículo publicado con el tema de evaluación referida a criterio, debido principalmente a que es el primero en plantear lo inadecuado de sustentar una variedad de decisiones acerca de programas instruccionales basados en objetivos o en competencias, a partir de mediciones obtenidas con las tradicionales pruebas normativas; el mismo Glaser, no obstante, señala que Flanagan en 1951 y Ebel en 1962 habían hecho una distinción entre evaluación normativa y criterial (Glaser, 1963). A partir de mediados de los años sesenta, se pueden encontrar en la literatura diversas definiciones de lo que es una prueba referida a criterio, entre las que destacan, por sus contribuciones al campo, la de Glaser y Nitko (1971), quienes enfatizan que una prueba criterial sirve para obtener mediciones directamente interpretables en términos de realizaciones estándar concretas, es decir, lo que el sustentante puede o no realizar; y la de Popham (1978), quien refería el uso de estas pruebas para determinar la posición de un individuo con respecto a un dominio perfectamente definido.

En cuanto a su uso, Hambleton y Swaminathan (1978) señalan que las evaluaciones referidas a criterio sirven para guiar el proceso individual en programas basados en objetivos de aprendizaje, comprobar el rendimiento de los alumnos, diagnosticar deficiencias de aprendizaje, evaluar programas educativos y de acción social y para verificar el logro de competencias con fines de certificación u otorgamiento de licencias. En este tipo de evaluación, la apreciación del grado con que un sustentante cumple con los objetivos de la enseñanza se lleva a cabo en función de su desempeño, sin compararle con el de sus compañeros.

Aunque existen algunas variantes en la definición del concepto de evaluación criterial,

hay acuerdo en que sirve para comprobar el rendimiento mediante la apreciación de las realizaciones personales respecto de los objetivos logrados, sin compararlas con las del grupo al que pertenece, facilitando así el diagnóstico de dificultades, la programación de las actividades de recuperación y la toma de decisiones de promoción de nivel o de certificación de cada individuo evaluado.

Otros autores importantes que aportaron herramientas conceptuales al campo fueron Popham y Husek (1969), con una aclaración importante al destacar que el término *criterio* —de las pruebas criteriales— se refiere a un dominio de conductas bien definido y no sólo al hecho de haber establecido un estándar de ejecución o un punto de corte. En cuanto a la interpretación de la ejecución o desempeño de un individuo, es correcto hacer una interpretación de tipo descriptivo, o bien establecer algún estándar o punto de corte contra el cual contrastar dicha ejecución, lo cual será posible siempre que se haya definido adecuadamente el marco lógico contra el cual se va a contrastar tal ejecución o desempeño, es decir, que la adecuada definición del dominio resulta indispensable. Desde luego la decisión de establecer un punto de corte o hacer una interpretación descriptiva depende del propósito de la prueba, como se puede apreciar en las distintas pruebas que se aplican en la actualidad.

Una confusión frecuente en este campo es la relativa a las diversas denominaciones que se han empleado en la literatura para referirse a pruebas que presentan pocas o ninguna diferencia con las pruebas referidas a criterio (las denominadas *pruebas de maestría*, o las pruebas *referidas a dominio*, y las *referidas a objetivos*). Si se adopta la definición de Popham (1978), no existen diferencias esenciales entre ellas; finalmente todas se constituyen de reactivos que emparejan con objetivos. La distinción principal entre pruebas *referidas a objetivos* y las *referidas a criterio* es que en estas últimas los ítems son un conjunto representativo de un dominio claramente definido de conductas que miden un objetivo, mientras que en las pruebas referidas a objetivos no siempre se especifica un dominio de conductas, por lo que no se considera que los ítems sean representativos de algún dominio conductual (Hambleton y Swaminathan, 1978).

Actualmente existen pocas dudas respecto de la necesidad de emplear pruebas con especificaciones cualitativamente diferentes a aquéllas que típicamente se han empleado para el diseño de una prueba referida a norma, y existe acuerdo en que las características más destacadas y comunes a lo que se ha dado en considerar una prueba referida a criterio son:

1. Requiere la definición clara y exhaustiva de un dominio objetivo a evaluar.
2. Permite averiguar la posición de un sujeto respecto del dominio de una conducta bien definida que manifieste el aprendizaje de un alumno.
3. La interpretación del rendimiento es directa: la ejecución que realiza el alumno indica su grado de dominio o competencia, independientemente de lo que hagan otros sujetos.
4. El criterio o estándar en el cual se basa tiene un carácter absoluto, es decir que no está condicionado por el nivel de ejecución de un grupo. Es la descripción de la clase de conducta que el alumno puede o no manifestar.
5. El límite en que se basa la toma de decisiones que afectan al proceso educativo se establece de manera descriptiva, indicando el grado de dominio alcanzado o bien especificando un punto que se toma como *punto de corte*, o *nivel mínimo de dominio.*
6. Permite retroalimentar la intervención en el proceso educativo de manera inmediata.

En términos metodológicos las pruebas referidas a criterio son especialmente útiles cuando se trabaja con unidades de aprendizaje

muy concretas, es decir, aquéllas al servicio de la evaluación formativa (Jornet y Suárez, 1994); no obstante, en este artículo se destaca su utilidad para la evaluación sumativa aplicada a cursos escolares completos o incluso a periodos más largos aplicada a grandes poblaciones, como es el caso de las pruebas a gran escala que se están aplicando actualmente en todo el mundo. Esta utilidad radica en su capacidad para:

1. Determinar el nivel de aprendizaje alcanzado por un alumno y, en función del mismo, asignarle una calificación o emitir un juicio de acreditación académica.
2. Certificar que un determinado alumno posee, al menos en el momento de la evaluación, ciertas destrezas, conocimientos, habilidades o competencias.
3. Predecir el éxito de un alumno en un curso posterior relacionado con la misma materia. Para que dicha predicción quede garantizada, el alumno al menos debe superar el punto de corte de aquellos objetivos que se consideran básicos, por su papel de conectores con los aprendizajes posteriores.
4. Indicar a los profesores, planeadores y autoridades educativas problemas de programación y secuenciación de áreas o temas en el currículo escolar.
5. Informar a los alumnos sobre su progreso en el aprendizaje, indicando los aspectos más logrados y los más deficientes.

Construcción de las pruebas criteriales a gran escala

Dado que las pruebas criteriales son construidas con la finalidad de apoyar generalizaciones válidas respecto del desempeño de los individuos en relación con dominios especificados de contenido o de conducta, Hambleton (1985) y Popham (1978) han enfatizado que los objetivos conductuales por sí mismos no son suficientes para una especificación detallada de los reactivos de la prueba. Las especificaciones de los reactivos, según Popham (1978), deben dividirse en cuatro partes: descripción de un objetivo conductual; dirección y ejemplos de los tipos de reactivos a incluir; acotación y limitaciones respecto de la amplitud y dificultad del contenido; y descripción del tipo de respuestas esperadas, así como de respuestas incorrectas. Con base en la investigación desarrollada en el campo durante más de dos décadas, Hambleton (1995) presenta una propuesta metodológica, que aún se considera vigente, de 12 pasos a seguir para la construcción de este tipo de pruebas:

1. Preparación y selección de las especificaciones del dominio o los objetivos que se pretenden evaluar.
2. Descripción clara y detallada de las especificaciones: los propósitos de la prueba, el tipo de formato de reactivos deseable, el número de reactivos de la prueba y las instrucciones para los redactores de reactivos.
3. Redacción de los reactivos para medir los objetivos incluidos en la prueba, o versiones de la prueba si se requiere la elaboración de formas paralelas.
4. Edición inicial de los reactivos de la prueba por los individuos que los redactan.
5. Evaluación sistemática y consistente de reactivos en los pasos 2 y 3 para determinar su congruencia con los objetivos y para determinar su representatividad.
6. Edición adicional de reactivos con base en los datos del paso 5, descartando aquéllos que no midan adecuadamente los objetivos que se pretenden medir.
7. Integración de versiones de la prueba con base en las especificaciones realizadas.

8. Empleo de métodos para determinar estándares (puntos de corte) que permitan interpretar la ejecución de los sustentantes.
9. Administración de la prueba bajo condiciones estandarizadas de aplicación.
10. Investigación y análisis para recopilar evidencias necesarias de confiabilidad y validez de la prueba.
11. Preparación de un manual técnico de la prueba.
12. Compilación de datos técnicos, tanto de los reactivos como de la prueba, para reforzar la validez en términos de las inferencias que se llevan a cabo y el tipo de usos en condiciones diversas y con diferentes poblaciones examinadas.

Resulta conveniente destacar que las operaciones fundamentales para la elaboración de este tipo de pruebas son: la especificación del dominio, el análisis de los reactivos, la determinación de estándares o puntos de corte, y la determinación de la confiabilidad y la validez; y que éstos son *procesos iterativos* hasta lograr niveles ajustados y satisfactorios. La definición del dominio o universo de medida es origen y referencia de todos los demás, por lo cual debe reunir características de precisión tales, que en todo momento permita saber si un reactivo pertenece o no a dicho dominio; el punto de corte o estándar se entiende como el punto de superación o nivel mínimo, es decir, el valor que, dentro de un continuo de medida sobre el que se puede situar la ejecución de un individuo, sirve para diferenciar el grado de suficiencia del grado de insuficiencia en el cumplimiento de objetivos o el logro de competencias.

Conceptualización del dominio educativo

La definición y estructuración del dominio es uno de los dos temas centrales sobre los cuales gira la construcción de una prueba criterial. En general existe coincidencia en la afirmación de que es la calidad de la definición del dominio lo que permite referir las puntuaciones individuales a criterios internos a la tarea, o en otras palabras, a criterios de *calidad* definidos como de mínima competencia. Este tema es crucial por su importancia en la determinación de la validez de contenido y de constructo, así como en otros conceptos de validez que se han sugerido para este tipo de pruebas y que se explicarán con mayor amplitud y profundidad en la cuarta sección de este artículo.

En relación con los elementos del dominio se pueden identificar tres niveles de definición: taxonómico, por objetivos y por ítems o reactivos. El nivel taxonómico puede desempeñar un doble papel: a) como orientación del análisis de acercamiento al dominio; y b) a partir de la comprobación empírica, como la síntesis de resultados genéricos descriptivos de la realidad. De esta forma se establece un proceso de retroalimentación teórico-empírico-teórico a través de los niveles inferiores de descripción, es decir, los objetivos y reactivos a partir de los cuales se podrán hacer inferencias acerca de procesos mentales involucrados para su solución.

De acuerdo con Millman (1974), es más útil contar con objetivos ampliados, los cuales define como un enunciado extenso de una meta educativa que proporciona abundantes especificaciones relativas a la situación de la prueba, las alternativas de respuesta y los criterios de adecuación de la respuesta correcta. Según este autor, la importancia de esta guía adicional a los objetivos conductuales es que ayuda a la definición apropiada del dominio de reactivos de la prueba; y es especialmente apropiada y útil si lo que se pretende es la evaluación de competencias.

A partir de estos objetivos ampliados es posible hacer un análisis de tareas y con ello una definición más precisa que consiste en la especificación de los ítems o reactivos que se van a construir, es decir que los reactivos serán las unidades mínimas de medición. Para una revisión de los procedimientos que se han propuesto y desarrollado en el marco de

la evaluación criterial se recomienda consultar algunos ejemplos como el de la especificación del dominio en cuatro pasos de Popham (1978) y la teoría de facetas de Berk (1978) para la definición de contenido para cada objetivo de interés; está también la transformación de ítems de Anderson (1972) y la de algoritmos de Scandura (1977). En un análisis de los métodos desarrollados, Hambleton (1980) señala que los mejores ejemplos de especificaciones de dominio eran los de Ebel (1962) y los de Hively, Patterson y Page (1968), quienes demostraron que es posible desarrollar y usar reglas de generación de reactivos para construir una prueba. Actualmente existen programas que utilizan estos desarrollos para la generación de pruebas educativas a gran escala, principalmente en el ámbito de la aplicación de pruebas adaptativas.

Jornet y Suárez (1994) presentan una conceptualización del dominio como universo de medida, a partir de la cual se analizan los elementos implicados en su definición y estructuración, y ofrecen una síntesis de los medios que han sido utilizados para hacer operativo el acceso y manipulación de dominios. Los autores aportan una visión genérica del dominio educativo que sirve para acercarse al problema de la medición desde una óptica más consecuente con la realidad educativa, compleja y multidimensional. Para ellos un dominio está bien definido si están especificadas sus unidades —objetivos y reactivos—, es decir que la calidad de la definición del dominio depende de la concreción de las unidades que lo definen. En este sentido, la definición de dominio se entiende en un continuo generalidad-concreción, de manera que a mayor concreción existe una mayor calidad en la definición. Como aspectos vinculados con la definición están los de exclusividad y exhaustividad, es decir que las unidades que definen el dominio no deben traslaparse y deben contener el dominio en su totalidad.

Lo anterior pone de manifiesto la importancia del material instruccional en cuestión, ya que a mayor ambigüedad en los contenidos, mayor será la generalización de la definición del dominio. Sólo si se conoce bien cuáles son los componentes y los procesos que subyacen a éstos, el dominio puede especificarse y medirse adecuadamente; cuando esto no es posible, se afectan directamente todas las características e indicadores integrados en el proceso de medición. Además de la definición, Jornet y Suárez (1994) destacan la importancia de analizar y especificar la estructura del dominio, para lo cual refieren dos sistemas de configuración: el de estructura implícita, que corresponde a aquellos dominios cuyas unidades tienen una característica propia independientemente de las demás unidades, como la dificultad teórica o complejidad cognitiva; y el de aquéllos cuya estructura es consecuencia de las relaciones entre las unidades, tales como relevancia del contenido, nivel de generalidad, secuencia, etc. El de estructura resultante corresponde, en cambio, a aquellos dominios obtenidos del análisis empírico de las respuestas de los sujetos a los reactivos, como lo es la dificultad.

La complejidad que implica la definición del dominio, así como su adecuada representatividad mediante un buen esquema de muestreo, requiere de una metodología para la elaboración de las especificaciones necesarias que garanticen la validez de constructo de la prueba. Tal metodología implica, además de la integración de profesores especialistas en el tema a evaluar, el uso de alguna taxonomía de objetivos como la que publicó Bloom desde 1956 y que fue ampliamente utilizada y difundida para este propósito durante más de medio siglo. Conviene más, sin embargo, utilizar una más actual, que haya incorporado los descubrimientos más recientes en torno a la representación y uso del conocimiento derivados tanto de investigación básica de las ciencias cognoscitivas, como de la investigación aplicada en educación. Una de ellas es la taxonomía de Marzano (2007), que además de integrar todos estos avances, inicia con una

crítica sustentada a la taxonomía de Bloom recuperando lo que aún es vigente, lo cual la hace apropiada para el enfoque actual de evaluación de competencias. Este hecho facilita su uso y comprensión por los profesores y especialistas, quienes necesariamente han de implicarse en la construcción de este tipo de pruebas y están familiarizados con la taxonomía de Bloom.

Otra taxonomía interesante es la propuesta por Biggs y Collins (1982), la cual se refiere al sistema de categorías y progreso jerárquico en la complejidad estructural de las respuestas de los escolares en el que expresan el aprendizaje (*Structured of the Observed Learning Outcome*, SOLO, por sus siglas en inglés). Esta taxonomía permite clasificar y evaluar el resultado de una tarea de aprendizaje en función de su organización estructural, ya que cada uno de los niveles describe un desempeño particular en un determinado momento, el cual se basa en el progreso de los estudiantes en tareas con niveles de complejidad creciente, mediante la relación de sus respuestas con aspectos más abstractos de las tareas, es decir que el proceso pasa de un conocimiento pobre o superficial a un conocimiento sólido o profundo de la realidad.

Cualquiera que sea la taxonomía elegida, es importante conocer sus fundamentos teóricos y garantizar su comprensión por parte de los especialistas, con la finalidad de evitar sesgos en la definición del dominio.

Determinación de estándares y puntos de corte

El segundo elemento central en la construcción de pruebas criteriales es el establecimiento de estándares o puntos de corte, ya que entre sus principales propósitos está la toma de decisiones que conciernen al control y seguimiento del progreso de los estudiantes a través del currículo para la promoción, certificación y graduación. Un estándar es un punto en la escala de puntuaciones de una prueba que sirve para clasificar, a quienes fueron examinados, en categorías que reflejan diferentes niveles de ejecución en relación con los objetivos o competencias medidos por la prueba. Existe una gran diversidad de métodos para establecer estándares y puntos de corte; para elegir el más adecuado debemos considerar algunos factores tales como: la importancia de las decisiones que se tomarán; el tiempo, las fuentes y recursos disponibles; la capacidad de los jueces especialistas con los que contamos (algunos métodos requieren mayor conocimiento del dominio del contenido, mientras que otros requieren mayor conocimiento de los estudiantes que serán examinados), y la pertinencia del método para el tipo de prueba que estamos elaborando. En cualquier caso, Cizek y Bunch (2007) apuntan reiteradamente la importancia de que las personas involucradas en la determinación de estándares o puntos de corte sean las mismas que participen en la definición del dominio, y cuando esto no sea posible, insisten en que al menos debe existir un puente de comunicación entre unos y otros especialistas.

En un intento por sintetizar la información relativa a los diversos métodos disponibles y ofrecer un panorama general de ellos, es importante destacar una diferencia entre procedimientos orientados a determinar el punto de corte a partir de una escala de puntuaciones verdaderas, de aquéllos que lo establecen a partir de puntuaciones observadas. En el primer caso, de acuerdo con Jornet y Suárez (1987), se refiere al establecimiento de un estándar, el cual se reservaría para hacer referencia al sistema de criterios de interpretación, la definición teórica de los niveles de desempeño, logro o competencia. De modo general los procedimientos utilizan el juicio de grupos de jueces expertos acerca de la prueba, de los individuos o de grupos de individuos. Por otra parte, los procedimientos utilizados para establecer un punto de corte a partir de un valor estándar que suponen previamente determinado, se ocupan de trasladar dicho valor a la escala observada, tomando en cuenta las diferencias que se dan en

la medición y generalmente optimizando las consecuencias resultantes de la decisión. En el Cuadro 1 se presentan los métodos en función del tipo de objetivo (Hambleton, 1980).

Cuadro 1. Métodos para la determinación de punto de corte

Tipo de objetivo	Categorías metodológicas
Determinación de un estándar	Métodos basados en juicios sobre los ítems
	Métodos basados en juicios sobre los sujetos
	Métodos basados en juicios sobre el grupo de referencia
Establecimiento de un punto de corte	Modelos de estado
	Modelos continuos basados en la teoría de la decisión
	Métodos de compromiso
	Otros métodos

Para la determinación de estándares, Linn (1979) propone una tipología en función del tipo de estándar: de exhortación, cuando representa metas deseables de logro a las que debe orientarse la mejora de un sistema educativo o de los estudiantes; de ejemplificación, cuando lo que representa son las competencias características de diversos niveles de ejecución; y de rendición de cuentas, cuando representan metas curriculares precisas, orientando la evaluación hacia la contrastación entre el currículum diseñado, el implementado y los logros obtenidos por los examinados.

En términos de los métodos para la determinación de puntos de corte, en el Cuadro 2 se ofrece una clasificación de los métodos más usados en el campo de la evaluación criterial en los inicios de las aplicaciones de este tipo de pruebas.

Cuadro 2. Clasificación de los métodos de establecimiento de puntos de corte

Métodos de juicio	Métodos empíricos	Combinación de métodos (mixtos)
Contenido del reactivo Nedelsky (1954) Angoff (1971) Angoff modificado (ETS, 1976) Ebel (1979) Jaeger (1978)	Criterio de medición Livingston (1975) Livingston (1976) Van der Linden y Mellenbergh (1977)	Juicios-empírico Grupos de contraste Zieky y Livingston (1977) Grupos en el límite Zieky y Livingston (1977) Grupos de criterio Berk (1976)
Adivinación Millman (1973)	Decisión teorética Kriewall (1972)	Consecuencias educativas Block (1972)
		Métodos bayesianos Hambleton y Novick (1973) Schoon, Gullion y Ferrara (1978)

Fuente: Hambleton, 1980: 104.

Los métodos llamados de juicio son los que mantienen el objetivo original de determinar un estándar absoluto capaz de diferenciar el nivel mínimo que debe satisfacerse, sin embargo, su uso exclusivo genera diversos problemas con su aplicación. Las críticas se han centrado principalmente en los métodos referidos al contenido de la prueba o juicio

sobre los ítems y son principalmente dos: 1) las diferencias encontradas entre los estándares producidos por métodos diferentes y 2) las discrepancias observadas entre los jueces dentro de un mismo método. Para mayor información sobre estudios comparativos entre métodos se recomienda el artículo de Jornet y Suárez (1987). No obstante, las discrepancias observadas no invalidan los métodos, dado que reflejan las diferencias predecibles a partir de cómo define cada método el nivel mínimo de competencia.

Hay autores que se han abocado a clarificar cuáles son los motivos conceptuales y técnicos que generan estas discrepancias entre los estándares resultantes de cada método, como es el caso de Brennan y Cockwood (1980) y Shepard (1980). Sin embargo, persiste el principal problema con los comités o grupos de expertos que establecen estándares únicamente mediante métodos de juicio, y es que a menudo establecen estándares que harían fallar a más de la mitad de los individuos evaluados, aún aquéllos que han completado todos los programas educativos acreditados y cuentan con experiencia práctica bajo estrecha supervisión (Schoon, Guillion y Ferrara, 1979). Una experiencia semejante a la de estos autores se vivió en México durante la etapa en la que los consejos técnicos enfrentaron la tarea de establecer los puntos de corte de los Exámenes Generales para el Egreso de la Licenciatura (EGEL), lo que condujo a evitar que se utilizaran métodos atendiendo sólo al contenido de la prueba; en esa ocasión se mostró a los integrantes de los consejos técnicos evidencia empírica de cuántos sustentantes no superarían tales estándares a pesar de haber acreditado todas las asignaturas de una licenciatura. Lo anterior deja un margen de duda acerca de qué se está evaluando, si el nivel mínimo de competencia real o lo que los jueces creen que debería ser (Leyva, 2004). En este tipo de pruebas, y dadas las implicaciones políticas que tienen, conviene que los estándares se definan por los jueces desde el momento en que se plantea la prueba, pero que se verifiquen y ajusten a partir de los resultados en una o varias aplicaciones con la población seleccionada.

Estudios comparativos de métodos de establecimiento de estándares

Toda esta polémica, y las discrepancias observadas, hicieron que durante la década de los ochenta se realizaran estudios de comparación de puntos de corte producidos por diversos métodos, de los cuales más de la mitad se dedicaron a comparar los métodos de juicio, principalmente los de Angoff, Ebel y Nedelsky; los estudios restantes trataron con uno o dos de estos métodos de juicio y los métodos de contraste y de límite (Livingston y Zieky, 1982). Entre los resultados más interesantes de estos estudios está la subjetividad de las decisiones empleadas a partir del contenido de los reactivos de los métodos de juicio, ya que los métodos producen estándares marcadamente diferentes cuando se aplica la misma prueba, ya sea por los mismos jueces o por muestras de jueces paralelas al azar (Shepard, 1980). Otros datos empíricos demuestran que los métodos de Angoff y Nedelsky tienen serios problemas debido a inconsistencias en las especificaciones de probabilidades de éxito (Van der Linden, 1984); no obstante, los autores señalan que es posible mejorar la exactitud de los métodos mediante la confrontación del estándar con los resultados inmediatos. Entre los métodos contrastados algunos demostraron ser más efectivos que otros para determinados casos; por ejemplo, en aquellos casos en los que era requerido un método de juicio para el establecimiento de estándares, el método de Angoff ofreció un buen balance entre adecuación técnica y practicidad.

En cuanto a los métodos empíricos, el punto de corte es elegido de manera sistemática a partir de cómo se distribuyen los resultados de los examinados; no obstante, los principales estudiosos del tema sugieren no usar sólo datos empíricos, ya que parte esencial de las pruebas criteriales requiere del componente de juicio

respecto a criterios o estándares logrados; en vez de ello recomiendan el uso de métodos mixtos, es decir, combinan los juicios de expertos y la evidencia empírica de su aplicación incluyendo datos de ejecución real en los procesos de establecimiento del punto de corte. El que toma las decisiones debe asignar peso, de manera primaria, a la evidencia de juicios.

Dado que una solución empírica para el problema del establecimiento de los puntos de corte no necesariamente resuelve el problema, no debe subestimarse el papel de los juicios en los métodos de grupos de contraste y grupos de criterio. Los juicios acerca de personas examinadas proporcionan fundamentos para la estimación estadística de las probabilidades de clasificación (Berk, 1996). El componente de juicio en estos métodos consiste en definir, operacionalmente, maestría o competencia en términos de la ejecución real en la prueba de individuos que han sido juzgados como competentes por sus profesores, supervisores inmediatos o personas similares, aptas o capacitadas dentro de un dominio de habilidades, conocimientos o competencias similares a las evaluadas. Aún cuando exista rigor en la especificación del criterio de selección y en los métodos estandarizados empleados, la debilidad de estos métodos estriba justo en nominaciones tales como: calificado, apto o competente, así como en los procesos para identificar personas competentes o no competentes para incluirlos en los grupos de criterio. Las interpretaciones de competencia o maestría a partir de una lista bien definida de habilidades pueden ser diversas y comparativamente limitadas. De este grupo de métodos, los de grupos de contraste se perciben como los más apropiados, en términos de la adecuación técnica, respecto de los restantes métodos analizados, seguidos de los métodos de grupo criterio. Sin embargo, su principal utilidad radica en que constituyen métodos de validación de estándares y de los niveles de desempeño que producen, más que métodos para el establecimiento de puntos corte.

Tendencias actuales en el establecimiento de estándares

Aunque la mayor parte de los métodos desarrollados hasta la década de los ochenta no se utilizan en la actualidad tal y como fueron propuestos, algunas de sus modificaciones han sido muy exitosas y han dado origen a una nueva generación de métodos. La evolución de estos métodos ocurrió en términos de la orientación general de procesos, entre ellos las técnicas de trabajo con los jueces y los indicadores de convergencia de los juicios. Esto se dio como resultado de estudios comparativos entre métodos en diversos tipos de pruebas con propósitos y ámbitos de aplicación también diversos, los cuales permitieron analizar y detectar mejores prácticas en términos de pertinencia y practicidad, aspectos que confieren mayor madurez metodológica al campo.

También ocurrió un cambio significativo en las interpretaciones: se pasó de las dicotómicas para admisión o certificación, a las interpretaciones a partir de series graduadas de niveles de desempeño de las pruebas a gran escala para la evaluación de sistemas educativos, como es el caso del NAEP (*Nacional Assessment Educational Program*), que utiliza tres niveles de desempeño: básico, competente y avanzado (Cizek y Bunch, 2007). Otro ejemplo es el caso de España: Jornet y González (2009) refieren que el estudio de Diagnóstico del Sistema Educativo Estatal Español de 1998, identifica niveles de competencia a partir de los ítems característicos de cada uno de ellos, considerando su comportamiento empírico; en el ámbito internacional está el proyecto PISA, que ha adoptado un sistema politómico de cuatro niveles descriptivos para informar de sus resultados.

Dentro de esta evolución, el componente que se afianza y adquiere un mayor reconocimiento en cualquier método es el consenso intersubjetivo de los especialistas en el dominio a evaluar. Tanto las categorías de contenido como las descripciones de lo que los sujetos

evaluados son capaces de realizar en cada nivel de desempeño y la selección de los ítems característicos de cada uno de estos niveles, son componentes que se desarrollan a través de procesos de juicios de expertos. Jornet y González (2009) presentan un análisis de la evolución de los enfoques más recientes sobre esta problemática, de las aproximaciones para definir categorías de contenido en el desarrollo de estándares, y de los tipos de métodos para identificar puntos de corte; así como criterios que pueden apoyar en la elección del método de determinación de estándares. Estos autores destacan la importancia del consenso intersubjetivo como referencia precisa para el diseño y como garantía de calidad de los estándares. En otro artículo (Jornet, González y Suárez, 2010) presentan un estado del arte de los métodos para desarrollar procesos de validación de la determinación de estándares en pruebas de rendimiento educativo.

En una revisión más exhaustiva de métodos de nueva generación, Cizek y Bunch (2007) sugieren una clasificación en términos de las variantes en los procedimientos empleados y en el tipo de información utilizada, presentando las siguientes agrupaciones generales:

1. Métodos de consenso directo. El más representativo de este grupo de métodos es una alternativa a los métodos de Angoff y Nedelsky (Sireci, Hambleton y Pitoniak, 2004), el cual, además de ocupar significativamente menos tiempo de los jueces expertos, supera algunas de las críticas más usuales a los métodos de juicio, incorporando estrategias que permiten a los especialistas expresar sus opiniones para la colocación del punto de corte de forma directa sobre una escala, evitando el procedimiento de emitir juicios por cada uno de los reactivos de una prueba, como en los métodos tradicionales.
2. Los denominados métodos holistas, cuya característica principal es la evaluación de muestras completas del trabajo de un examinado por uno o varios jueces que rinden un solo dictamen global acerca de cada muestra de trabajo. La valoración tiene el propósito de clasificar los trabajos en categorías de rendimiento, o bien en categorías que representan los límites entre los niveles de rendimiento. Estos métodos son muy útiles para tareas de desarrollo o ejecución.
3. Los métodos que entran en esta clasificación son el del juicio analítico (Plake, Hambleton y Jaeger, 1997) y el del cuerpo de trabajo (*Body of work method*) (Kingstone *et al*., 2001), donde el juicio está basado en el examen de las respuestas en un amplio cuerpo de trabajo del estudiante, método que registra una mayor frecuencia de uso.
4. Métodos de correspondencia de ítems, como el método del marcador de Lewis, Mitzel y Green (1996), los cuales son vistos como una sucesión lógica de una serie de estrategias desarrolladas en los noventa en conjunción con el establecimiento de puntos de corte utilizados en el National Assessment of Educational Progress (NAEP) por investigadores del American College Testing (ACT) referidos como estimación media y que en esencia representan una extensión de la técnica modificada de Angoff. El procedimiento es un conjunto completo de actividades diseñadas para producir puntos de corte con base en la identificación, por parte de los jueces, de ítems que actúan como punto de inflexión entre dos niveles de desempeño previamente definidos por juicio. Los ítems están ordenados en términos de su dificultad empírica, lo que permite ajustes más realistas. Una variante de este método se empleó en el INEE como modelo de determinación de

niveles de logro de los EXCALE (Jornet y Backhoff, 2008).

5. Métodos de empate reactivo-descriptor, que consisten en describir el conocimiento y las habilidades esperadas de examinados en cada uno de los niveles de ejecución alcanzados, es decir, lo que son capaces de hacer por cada categoría de ejecución. Estos métodos comparten características con el método del marcador, tales como el uso de los conjuntos de reactivos ordenados por índice de dificultad; de hecho ambos métodos pueden considerarse como casos especiales de aproximaciones de mapeo de reactivos, sólo que el de empate ítem-descriptor se centra en los juicios de los jueces en áreas de incertidumbre de la clasificación, y en términos de los procedimientos analíticos empleados hay mucha similitud con los que se usan en los métodos de contraste.
6. Métodos de compromiso, que recomiendan mezclar información normativa y criterial. El punto de corte se establece mediante un acuerdo entre los niveles mínimos de competencia estimados por jueces y la distribución empírica resultante de un grupo de referencia. En esta categoría están los métodos de contraste y los más representativos son el método de compromiso de Hofstee (1983) y el de Beuk (1984).
7. Los métodos empíricos son de diversa índole y lo que tienen en común es el hecho de que la mayor parte del procedimiento se sustenta en información empírica; es decir que se caracterizan por utilizar la escala de puntuaciones observadas. De estos métodos podemos identificar tres grupos: a) los modelos de estado; b) los modelos continuos basados en la teoría de la decisión; y c) los modelos continuos basados en la distribución de los ítems sobre la escala de habilidad total.

En la actualidad se han venido planteado nuevos problemas e implicaciones importantes para la determinación de estándares y puntos de corte, como es el hecho de medir apropiadamente el progreso anual respecto de los estándares establecidos por grado, lo cual implica crear un sistema coherente de estándares de ejecución a través de los grados escolares y de los individuos que permita hacer inferencias acerca de si los estudiantes superaron los estándares de cada evaluación y del progreso anual que van logrando, de una manera significativa y lo más exacta posible (Cizek, 2005).

Una fuente de información más completa para apoyar las decisiones relativas al establecimiento de estándares o puntos de corte es la que publican Cizek y Bunch (2007), quienes además de hacer una reseña muy completa, describen cambios y futuras direcciones en el establecimiento de puntos de corte a partir de experiencias que ellos tuvieron tanto en el campo de pruebas a gran escala de certificación como en programas de evaluación formativa. Desarrollaron este trabajo en atención a los nuevos requerimientos de la legislación federal de los Estados Unidos de Norteamérica de administrar pruebas de tercero a octavo grado para medir el progreso anual de los estudiantes en lectura y matemáticas con la finalidad de proporcionar información útil a profesores, padres de familia y estudiantes.

Conocer estos métodos para establecer sistemas de estándares que sean aplicables a pruebas distintas para evaluar el progreso académico de los estudiantes resulta muy alentador en términos de la oportunidad que brindan para sustentar decisiones orientadas a elevar la calidad de la educación.

Confiabilidad y errores de medida

La utilidad de las mediciones de desempeño presupone que los individuos y los grupos exhiben algún grado de estabilidad o regularidad en su conducta. No obstante, muestras sucesivas de desempeño de una misma persona son

raramente idénticas en todos los aspectos; las ejecuciones, actitudes, productos y respuestas a conjuntos de preguntas varían en su calidad y carácter de una ocasión a otra, aún dentro de condiciones estrictamente controladas. Esta variación se refleja en las medidas que se obtienen mediante una prueba, y las causas de esta variabilidad generalmente no están relacionadas con los propósitos de la medición.

Un examinado puede esforzarse mucho durante una aplicación, o tener más suerte, o estar más alerta, o sentir menos ansiedad, o gozar de mejor salud en una ocasión que en otra. También puede ser que un examinado pueda tener mayor conocimiento, experiencia o comprensión de lo que es relevante a la tarea en el dominio muestreado en la prueba, que los demás examinados. Algunos individuos pueden exhibir menos variación en sus mediciones que otros, pero ninguna persona es completamente consistente, lo cual implica que siempre que se haga una medición existirá un error de medida derivado de alguna de las fuentes de variación antes señaladas.

La confiabilidad se refiere al grado en el cual las medidas de una prueba o de un procedimiento de medición están libres de error; es el grado de consistencia de tales mediciones cuando el procedimiento o la prueba son repetidos en una población de individuos o grupos (AERA, APA, NCME, 1999). Así como toda medida incluye un componente de error, existe un valor hipotético libre de error que caracteriza a un examinado en algún atributo o dominio representado en una prueba. En términos de la *teoría clásica* de las pruebas, este valor libre de error es la *medida verdadera* de la persona y se define como la medida promedio resultante de repeticiones de la prueba o de formas alternas del instrumento. En términos estadísticos es un parámetro personal, y cada medida observada es una estimación de este parámetro.

Dentro de la aproximación de la *teoría de la generalizabilidad*, un concepto comparable es referido como la *medida universo* del examinado; y dentro de la *teoría de respuesta al ítem (*IRT por sus siglas en inglés), un concepto similar es el llamado la *habilidad* de la persona. La diferencia hipotética entre una medida de la persona observada a través de un procedimiento de medición y su *habilidad* o *medida universo* o *medida verdadera* es lo que entendemos como el error de medida.

Los errores de medida son usualmente vistos como aleatorios e impredecibles, contrario a aquellos errores sistemáticos, los cuales pueden afectar la ejecución de individuos o grupos pero de una manera consistente, como por ejemplo aquéllos que resultan de la aplicación de formas alternas de una prueba que no son equivalentes, ya que las personas que tomen la forma más difícil tendrán una medida promedio menor que aquéllos que tomen la otra; en este caso, tal diferencia no debe considerarse como un error de medida dentro de los métodos de cuantificación y resumen del error. Los estándares de calidad señalan la importancia de la estandarización de pruebas y procedimientos para asegurar la consistencia en las principales características de las pruebas, así como en el apego a los procedimientos estipulados en la administración y el uso prescrito de las medidas obtenidas para reducir el error (AERA, APA, NCME, 1999). En el caso de la aproximación de la *teoría de la generalizabilidad* (TG), estas diferencias pueden reconocerse como una fuente de error.

El error de medida reduce la utilidad de la medición y limita el grado en el cual los resultados de una prueba pueden generalizarse más allá de las condiciones específicas de aplicación de la medición; no obstante, dada la naturaleza aleatoria del error de medida, no es posible separarlas de las medidas observadas, sólo es posible saber su magnitud a través de algunos procedimientos estadísticos. La información crítica en confiabilidad incluye la identificación de las principales fuentes de error, un resumen estadístico apoyado en la magnitud de tales errores y el grado de *generalizabilidad* de las medidas a través de formas

alternas, mediciones, administraciones o cualquier otra dimensión relevante.

La información acerca de la confiabilidad puede reportarse en términos de varianza de errores de medida, en términos de uno o más coeficientes, o en términos de *funciones de información de pruebas* basadas en IRT. El error estándar de medida es la desviación estándar de una distribución hipotética de errores de medida de una población evaluada mediante una prueba o procedimiento particular. Tradicionalmente se reconocen tres categorías de coeficientes de confiabilidad:

1. Coeficientes derivados de la administración de formas paralelas en sesiones de prueba independientes.
2. Coeficientes obtenidos por la administración del mismo instrumento en ocasiones separadas.
3. Coeficientes basados en la relación entre medidas derivadas de reactivos individuales o subconjuntos de reactivos dentro de una prueba, con datos de una misma aplicación denominados "coeficientes de consistencia interna".

Con el desarrollo de la TG las tres categorías mencionadas se consideran como casos especiales de una clasificación más general de coeficientes de *generalizabilidad*, los cuales son definidos, igual que los coeficientes tradicionales, como la razón entre la varianza de medidas verdaderas y la varianza de medidas observadas, pero con la salvedad de que permiten al investigador especificar y estimar los diversos componentes de varianza verdadera, varianza de error y varianza de medidas observadas mediante la aplicación de las técnicas de análisis de varianza (Kieffer, 1999). De especial interés son los estimados numéricos por separado de los componentes de varianza del error total, ya que permiten examinar la contribución de cada fuente de error, además de que hacen posible la estimación de coeficientes de confiabilidad aplicables a una amplia variedad de diseños de medición.

Una aportación de la IRT al campo de la evaluación mediante pruebas a gran escala son las funciones de información de pruebas, las cuales resumen eficientemente qué tan bien discrimina una prueba entre los diversos niveles de habilidad de los individuos; dentro de esta teoría se emplea una función matemática denominada la *curva característica del ítem* o *función de respuesta al ítem*, como modelo para representar el incremento en la proporción de respuestas correctas a un ítem por grupos de niveles de habilidad progresivamente mayores en el rasgo o característica que se está midiendo (Embretson y Reise, 2000). Esta función puede tomarse como una expresión matemática de la precisión de medida en cada nivel del rasgo o dominio evaluado. Precisión, en el contexto de la IRT, es análoga al recíproco de la varianza de error condicional de la teoría clásica.

Aunque los coeficientes de confiabilidad tradicionales, y aquéllos derivados de las otras dos aproximaciones, parecieran ser intercambiables, en realidad conllevan formas diferentes de información. Un coeficiente puede proporcionar información desde una perspectiva amplia, mientras que otros sólo desde un ámbito más restringido. Un coeficiente puede reflejar error debido a inconsistencia en la medición y no reflejar la variación que caracteriza a pruebas sucesivas de ejecuciones o productos; o bien puede reflejar la consistencia interna del instrumento y fallar en reflejar los errores de medida asociados con cambios en los examinados. Por otra parte, los errores estándar de medida pueden reflejar variaciones de muchas fuentes de error o sólo de algunas, por lo que es necesario tener especial cuidado en la elección e interpretación de los diversos índices o coeficientes que se incluirán en un estudio de confiabilidad, así como la decisión de la magnitud y tipo de error que se puede aceptar, dependiendo del uso específico que se le dará a la prueba.

No es fácil recomendar o decidir entre las opciones de cuantificación de la confiabilidad; ningún método de investigación es óptimo en todos los casos, ni es recomendable limitarse a sólo una aproximación. Es por eso que los estándares de la AERA, APA y NCME (1999) demandan a los evaluadores que reporten no sólo los coeficientes de confiabilidad, sino el detalle de los métodos empleados para estimarlos, la naturaleza de los grupos o individuos de los que se derivan los datos, las condiciones dentro de las cuales fueron obtenidos y el uso que se dará a las mediciones. Finalmente, es necesario reconocer y enfatizar que el nivel de confiabilidad de las medidas de una prueba tiene implicaciones para la validez de la interpretación de las mismas.

La validez en las pruebas criteriales

La validez es la consideración fundamental en el desarrollo y evaluación de una prueba. El concepto se refiere al grado en el cual la evidencia y la teoría apoyan las interpretaciones de las medidas de una prueba de acuerdo con los usos previstos (AERA, APA, NCME, 1999). La validación de una prueba es el proceso de acumulación de evidencias que apoyen tales interpretaciones, y su objetivo es determinar qué tan apropiadas, significativas y útiles resultan las inferencias específicas que se hacen a partir de las mediciones realizadas mediante la prueba en función del uso específico para el cual se diseñó.

Se pueden identificar diversas fuentes de evidencia que pueden aclarar algunos aspectos de la validez; no obstante se trata de un concepto unitario, es el grado en el cual toda la evidencia acumulada apoya la interpretación de las medidas de una prueba de acuerdo con el propósito propuesto. Tradicionalmente han existido diversas maneras de obtener evidencias de validez, por lo que se han convenido categorías tales como la validez de contenido, de constructo y de criterio. Esta última puede ser de carácter predictivo o concurrente. Sin embargo, tales categorías y niveles no implican que existan distintos tipos de validez, o que alguna estrategia de validación sea mejor para cada tipo de inferencia o uso posible de una prueba; de hecho, no es posible hacer una distinción rigurosa entre ellas.

Un proceso de validación ideal incluye varios tipos de evidencia y, desde luego, la calidad de esta evidencia; una línea simple de evidencia sólida es preferible, en ocasiones, a numerosas y variadas líneas de evidencia cuya calidad sea cuestionable. Los juicios profesionales deben guiar las decisiones relativas a las formas de validación que son más necesarias y viables a la luz de la intención y el uso de la prueba. Los recursos deben dirigirse a obtener la combinación de evidencia que refleje de manera óptima el valor de la prueba para el propósito para el cual se construyó, por lo que el uso de diversas fuentes de información en los procesos de validación permite considerar aquellas variables y facetas importantes, obteniendo así una estimación más amplia que incluya también evidencia de la validez de clasificación de niveles de desempeño a partir de la determinación de estándares o puntos de corte.

En cuanto a las pruebas referidas a criterio, la validez debe estudiarse en relación con los usos principales de sus puntuaciones, según la propuesta de Hambleton (1984):

a) Describir lo que conocen los examinados en términos de ejecución.
b) Describir la ejecución de grupos específicos de sujetos en evaluación de programas.
c) Clasificar a los sujetos en niveles de desempeño.
d) Certificar la competencia de un individuo respecto de un dominio definido.

Messick (1975) comenta que lo que muchos evaluadores presentan como validez de contenido se basa únicamente en el análisis

formal de la congruencia reactivos-objetivo, siendo esto, más bien, relevancia de contenido o representatividad del contenido, pero no validez, ya que no proporciona evidencia que apoye la interpretación de respuestas o mediciones. En esta misma línea de ideas, Linn (1979) afirma que la cuestión de validez es una cuestión propia de la interpretación de la medida más que de la medida en sí. Las medidas pueden tener diversas interpretaciones, las cuales seguramente difieren en el grado de validez y por lo tanto en el tipo de evidencia que se requiere para el proceso de validación.

Lo más apropiado en la actualidad es conducir estudios de validez de constructo para validar el uso de las mediciones de la prueba, y en el caso de que la prueba se utilice para tomar decisiones respecto del nivel de competencia logrado en función de estándares o puntos de corte, se requiere validar tanto las clasificaciones que se producen como los mismos procedimientos o métodos mediante los cuales se establecieron dichos estándares.

Validez de contenido o evidencia basada en prueba de contenido

La calidad de los reactivos de una prueba se puede determinar por el grado en el cual éstos reflejan, en términos de su contenido, el dominio del cual se derivan. La evidencia de la validez basada en pruebas de contenido puede obtenerse mediante el análisis de la relación entre el contenido de la prueba y el constructo que intenta medir. La acumulación de evidencia involucra una consideración de tres características de los reactivos de una prueba: que el reactivo realmente mida algún aspecto del contenido incluido en las especificaciones del dominio, su calidad técnica y su representatividad. Es decir que se pretende establecer si la prueba es una muestra adecuada o representativa del dominio, y se favorece a partir de la calidad de la definición del dominio, de la propia calidad técnica de sus reactivos y del sistema de muestreo utilizado para construir la prueba.

El perfil referencial de validez y las tablas de especificaciones son los documentos necesarios para garantizar a priori la validez de contenido y de constructo de las pruebas. Las tablas de especificaciones son también instrumentos base para los procedimientos de validación por jueces, los cuales tienen que emitir juicios respecto de si los reactivos son adecuados y pertinentes al perfil referencial y a la definición del dominio que queda establecido en la tabla de especificaciones. Aún cuando se sigan todos los pasos descritos en las secciones anteriores para la definición y estructuración del dominio, las especificaciones no son siempre lo suficientemente precisas para asumir a priori que los reactivos que se generan son válidos, por lo que independientemente de lo cuidadoso que sea el proceso de generación de reactivos, se deben conducir estudios *a posteriori*.

Hay dos aproximaciones generales que se usan para establecer la validez de contenido de reactivos de una prueba referida a criterio: la primera aproximación involucra los juicios emitidos por especialistas en el contenido. Estos juicios conciernen al grado en que un reactivo es congruente y pertinente con el dominio que está destinado a medir. La segunda aproximación consiste en aplicar técnicas empíricas en la misma forma en que se aplican a los ítems de las tradicionales pruebas referidas a norma.

Al respecto, Hambleton (1980) propone algunos métodos derivados de la primera aproximación tanto para obtener evidencia de la validez de contenido como de la calidad técnica del reactivo; respecto de la representatividad señala que para poder determinarla se requiere integrar alguna versión de la prueba, y si el dominio está definido con claridad los especialistas podrán emitir sus juicios acerca de la representatividad de los reactivos. Lo cierto es que ambas aproximaciones son perspectivas complementarias, por lo que lo más adecuado es que se aborden tanto la *revisión lógica*, esencial para la selección de los

reactivos, como la *revisión empírica*, enfoque que complementa al primero y que permite su comprobación. La revisión empírica se orienta a la obtención de información acerca del funcionamiento de los ítems o reactivos y de la consistencia del funcionamiento de la prueba, y se concreta en el análisis de datos para comprobar las hipótesis de dificultad, discriminación, ajuste y validez en la interpretación de resultados.

Validez de constructo y elementos de estructuración del dominio

Desde un punto de vista científico, el término validez se refiere a la validez de constructo; mientras que los términos validez predictiva, concurrente, convergente, factorial etc., pueden ser considerados más bien como estrategias de colección y análisis de datos empleadas para probar las conexiones conceptuales entre la medición y el constructo (Angoff, 1988). La validez se entiende como la existencia de evidencias en torno a la consistencia entre el perfil referencial y la prueba; el énfasis está dado en sustentar el grado en que los puntajes en la prueba representan la medida de la característica o atributo psicológico que se supone evalúa la prueba; es decir, el constructo teórico. Para ello es necesario establecer procedimientos de revisión lógica de la adecuación, del análisis de su estructura interna y del análisis de la relación de la prueba con variables externas.

También se deben establecer análisis de constructo en rangos que justifiquen los niveles de desempeño sobre los cuales se establecen los puntos de corte. Tanto las descripciones como las decisiones que se toman a partir de una prueba referida a criterio se hacen con base en las respuestas que los sustentantes dan a los reactivos de la prueba, por lo que es esencial establecer un diseño experimental cuidadoso para investigar la validez de constructo. Estas investigaciones deben derivarse necesariamente del uso propuesto de las mediciones de la prueba, ya que éste proporcionará la dirección para el tipo de evidencia que es prioritario recuperar.

Mientras que en algunos casos el dominio de medida de una prueba puede ser el criterio de interés, en otros casos puede existir la intención explícita de generalizar más allá del dominio de ítems de la prueba. La necesidad de hacer inferencias en un dominio más amplio del que la prueba mide directamente, requiere de bases teóricas más profundas que vinculen la prueba y el criterio; esto es, la necesidad de obtener evidencias de validez de constructo. Haertel (1985) expone la conveniencia de concebir los resultados de aprendizaje como constructos, y a las pruebas referidas a criterio como medidas de estos constructos. Estos constructos se contrastan con otros "más naturales" derivados de la investigación psicológica. Como estrategia de evaluación sugiere la integración de teorías de procesos psicológicos y estructuras de memoria implicadas en estos constructos con descripciones de ejecuciones demostradas en diversos contextos, dentro y fuera del ámbito escolar.

En el ámbito educativo generalmente los atributos son definidos primariamente en términos de sus manifestaciones conductuales, y sólo de manera secundaria en términos de los procesos cognoscitivos y mecanismos de memoria subyacentes. Con ellos se pretende cubrir o relacionar un rango específico de situaciones y son menos estables en el tiempo que la mayoría de los constructos psicológicos. De acuerdo con Haertel (1985), no obstante las diferencias que existen entre las pruebas psicológicas y las educativas, la validez de constructo es tan importante para justificar las interpretaciones de las pruebas educativas criteriales como para las mediciones psicológicas, por lo que existe una considerable similitud en la lógica de validación e interpretación de estas dos formas de medición.

Las pruebas psicológicas son empleadas para hacer inferencias de constructos no observables, para lo cual las predicciones e inferencias se hacen a partir de conductas

manifiestas. Las pruebas educativas se emplean para determinar si los estudiantes se desempeñan adecuadamente en algún dominio de contenido específico. Frecuentemente es necesario y deseable intentar generalizar a un dominio más amplio de situaciones y tipos de respuesta de las que la prueba contiene, como es el caso actual de la evaluación de competencias; en este caso se deberá justificar dicha generalización por medio de alguna teoría psicológica (de aprendizaje, memoria, recuperación o transferencia).

La validez de constructo no sólo sirve para justificar los usos de una prueba educativa, sino que puede proporcionar una articulación entre líneas de investigación de la psicología educativa con la psicología cognoscitiva, colocando el énfasis en los procesos cognoscitivos y las estructuras de memoria desarrolladas mediante el proceso de instrucción (Greeno, 1980; Snow, 1980). Zeller (1988) propone seis pasos necesarios para establecer la validez de constructo:

1. Elegir o construir una teoría para la definición de conceptos y la determinación a priori de las relaciones entre ellos.
2. Seleccionar indicadores que representen cada uno de los conceptos contenidos en la teoría.
3. Establecer la naturaleza dimensional de estos indicadores.
4. Calcular la correlación entre las escalas construidas.
5. Comparar las correlaciones empíricas con las relaciones teóricamente determinadas entre los conceptos.

Queda claro que la evaluación o constatación de los dos primeros pasos requiere de análisis lógicos a través de jueces expertos, mientras que generalmente los últimos tres pasos se han llevado a cabo mediante análisis empíricos como el análisis factorial, técnica usualmente empleada para el análisis de dimensiones de una prueba referida a norma, ya que permite corroborar, a través de una matriz, si un patrón de factores obtenido en el análisis corresponde con el patrón de objetivos especificados en el dominio, y si cada reactivo forma parte del factor/objetivo predeterminado.

La estructura resultante de un análisis factorial se compara con alguna estructura que especifique una relación teórica entre los objetivos. Debe quedar claro que los procesos de validez de constructo están necesariamente ligados a la teoría, y por ello es materialmente imposible validar la medida de un atributo si no existe una red teórica subyacente al atributo a evaluar. El significado de un factor no depende de las características estadísticas de sus indicadores, sino de su contenido teórico. Para decidir adecuadamente cuál de los factores empíricos representa adecuadamente la estructura del dominio a evaluar, es necesario ir más allá de los criterios estadísticos usados.

Entre las técnicas que se reportan en la literatura como de uso más frecuente para determinar la validez de constructo de una prueba referida a criterio está la del análisis del escalograma de Guttman, siempre y cuando los objetivos puedan ordenarse en secuencias lineales o jerárquicas. En la medida en que las mediciones obtenidas respecto de un objetivo de la jerarquía concuerden con la jerarquía establecida se estará ofreciendo evidencia de validez de constructo; si, por el contrario, las mediciones no concuerdan, puede pensarse que ha ocurrido una de tres posibles situaciones: la jerarquía está incorrectamente especificada, las mediciones de los objetivos no son válidas, o una combinación de ambas explicaciones.

Validez de criterio o evidencia basada en relaciones con otras variables

Aún cuando las puntuaciones derivadas de las pruebas referidas a criterio sean descriptivas de los objetivos que suponen reflejar, la utilidad de éstas como predictores para decir

que un sustentante tendrá o no tendrá éxito en la siguiente unidad de instrucción no se puede asegurar. Los estudios de validez de criterio que se emplean en las pruebas normativas son los mismos que se pueden usar para las pruebas criteriales (Cronbach, 1971).

La validez criterial se entiende como la consistencia entre las decisiones que puedan derivarse a partir de la prueba y las de otro instrumento o proceso alternativo externo a la prueba que sirva como criterio para la misma. Se establece mediante estudios de validez concurrente —relación con otras pruebas o formas de evaluación con el mismo significado teórico— o validez predictiva, mediante la cual se analiza la capacidad de la prueba para cumplir los objetivos fijados en cuanto a su potencialidad para predecir acontecimientos ulteriores, como por ejemplo el rendimiento del alumnado en niveles o cursos posteriores. La evidencia obtenida mediante la relación con otras variables se orienta a determinar el grado en el cual estas relaciones son consistentes con el constructo de la prueba.

Otro tipo de estudios que entran dentro de esta categoría es la de grupos de contraste para establecer la validez de clasificación o decisión. Las pruebas criteriales comúnmente se emplean para tomar decisiones en donde se espera que la ejecución de un sustentante exceda un nivel mínimo de ejecución, a menudo referido a un estándar, para considerarlo apto, es decir, con un nivel de desempeño satisfactorio, generalmente para promoverlo u otorgarle algún certificado. El análisis de validez de decisión generalmente se realiza por medio de jueces expertos en el área, los cuales estudian las propiedades que engloban los reactivos ordenados en el modelo de Guttman. Las categorías no deben ser más numerosas que los niveles de desempeño que van a dictaminarse.

Este tipo de validez es en realidad una forma particular de validez de constructo, e involucra el conjunto de estándares de ejecución de una prueba y la comparación de la ejecución de la prueba de dos o más grupos de criterios en relación con el estándar especificado (Leyva, 2004). Los grupos se forman considerando algún criterio que determine su grado de maestría respecto del dominio a evaluar; por ejemplo, expertos *vs.* novatos. En este caso se aplica la prueba a ambos grupos y se obtiene el porcentaje de sujetos clasificados correctamente mediante la prueba. La ventaja de este procedimiento radica en que es reportada en una forma interpretable. Adicionalmente, la correlación entre dos variables dicotómicas (miembros de un grupo contra decisiones de maestría) puede reportarse y emplearse como índice de validez de decisión.

La validez de decisión depende de varios factores importantes: 1) la calidad de la investigación de la prueba; 2) la pertinencia de los grupos de criterio; 3) los grupos examinados; y 4) el nivel mínimo de ejecución requerida para alcanzar el nivel de maestría o competencia denominado estándar o punto de corte. Los puntos de corte deben ser validados en combinación con el análisis de reactivos, de tal forma que se garantice que los reactivos se ubican bien en un constructo dado, permitiendo ser usados como discriminadores de los niveles de desempeño, lo que en la literatura se denomina "anclaje" de los reactivos a los niveles de desempeño. Al respecto, Berk (1976) estima que el mejor punto de corte es aquel que maximice la validez de clasificación; la utilidad y la validez se incrementan minimizando los errores de clasificación. Adicionalmente, para las pruebas criteriales, parece de gran utilidad el método de validación de grupos de criterio (Berk, 1976). Se basa en la utilización de dos grupos de sujetos como criterios de contraste: los instruidos (que han superado con éxito un curso) y los no instruidos (que aún no lo han abordado).

Otra alternativa, cuando se han cuidado las propiedades métricas de la prueba, sería el anclaje de escala propuesto por Beaton y Allen (1992), el cual involucra un componente

estadístico que identifica reactivos que discriminan entre puntos sucesivos en una escala de ejecución usando características específicas de los reactivos. También involucra un componente de consenso en el cual se emplean reactivos identificados por expertos especialistas en el área, para proporcionar una interpretación de lo que saben o pueden hacer los grupos de estudiantes en, o cerca de las puntuaciones de las escalas seleccionadas.

Tanto las escalas referidas a norma como las referidas a criterio contienen información útil para quien trata de interpretar los resultados de una prueba. El *National Assessment of Educational Progress* (NAEP) intenta satisfacer ambos tipos de interpretaciones mediante la producción de escalas continuas que sean manejables para la interpretación referida a norma y por el anclaje de estas escalas en una forma que describa, en términos probabilísticos, lo que saben o saben hacer estudiantes de diferentes puntos de la escala. La idea básica del anclaje es simple, sin embargo, es probable que los intentos por describir los logros de los estudiantes en cada punto de la escala resulten complicados. Por ello es conveniente elegir algunos puntos a lo largo de la escala para la descripción, los cuales se denominarán "puntajes o niveles ancla" (Beaton y Allen, 1992). Es probable que en muchos casos el nivel de logro sea acumulativo, es decir que los estudiantes de mayor nivel de desempeño sepan y puedan realizar todo lo que saben y pueden hacer los estudiantes de niveles más bajos y más. Es por ello que las descripciones deben incorporar el incremento en el logro entre los diferentes puntajes ancla de la escala.

Hay dos métodos de anclaje de la escala: el método directo y el método atenuado. Ambos requieren que la escala haya sido generada por métodos psicométricos tradicionales o de teoría de respuesta al ítem (IRT). Los métodos directos usan las funciones discretas de respuesta a los reactivos, es decir, la proporción de respuestas correctas en los diferentes niveles de la escala. El método atenuado emplea el procedimiento de ajuste de curvas a los reactivos para crear funciones atenuadas de respuesta al ítem.

Validez cognoscitiva

Como ya se señaló, en el enfoque actual de evaluación de competencias se requiere expandir la teoría de validez hacia una teoría que dé cuenta de interpretaciones de procesos cognoscitivos a partir de medidas de una prueba o de ejecuciones observables para poder generalizar los resultados más allá del dominio de una prueba. Shavelson y Ruiz-Primo (2000) proponen algunas aplicaciones de métodos de evaluación tales como el mapa conceptual para establecer la validez cognoscitiva de las medidas de pruebas de ejecución o competencias. De acuerdo con esta aproximación, los resultados de una prueba constituyen una muestra del universo de conductas de interés del individuo, a partir de la cual se hacen inferencias.

Desde esta perspectiva, el puntaje o medida asignada a un estudiante es una muestra del dominio de posibles medidas que pudimos haber obtenido del estudiante. Un esquema de muestreo es útil para identificar facetas que caracterizan la medición. Las facetas incluyen: a) la tarea presentada, b) la ocasión de la medición, c) los juicios de quienes observan la ejecución, y d) los métodos de evaluación (Ruiz-Primo y Shavelson, 2001). Esto significa que, para un tipo particular de evaluación, las facetas relevantes a la medida pueden variar, por ejemplo, la faceta que se refiere a los juicios de observadores es irrelevante en una prueba de opción múltiple.

Tradicionalmente la variación debida a la tarea, la ocasión o los juicios, se manejaba como fuente que atentaba contra la confiabilidad de la prueba. En contraste, la incorporación de métodos de medición dentro de la especificación del universo en el cual ocurre el muestreo, nos permite trasladar nuestro enfoque de una teoría de confiabilidad a una teoría de validez. Cuando las ejecuciones

varían de una tarea a otra, o de una ocasión a otra, hablamos de un error de medida debido a la variabilidad del muestreo. Pero si la ejecución varía de un método de medición a otro hablamos de un problema de validez (convergente) debida a variabilidad ocasionada por el método (Kane, 1982; Baxter y Shavelson, 1994).

Este tipo de evidencia de validez de constructo se orienta a examinar las interpretaciones propuestas de los puntajes de una prueba mediante estudios de investigación que impliquen comparaciones entre expertos y novatos, análisis cognoscitivo y análisis de la calidad de las tareas (Shavelson y Ruiz-Primo, 2000; Ruiz-Primo y Shavelson, 2001). La importancia de realizar este tipo de investigación es que permite profundizar en los procesos cognoscitivos evocados con la solución de casos o problemas, proporcionando elementos valiosos para la construcción de programas educativos más adecuados para el desarrollo de competencias en los diferentes niveles educativos (Patel, Kaufman y Arocha, 2000). También proporciona elementos para mejorar los métodos de evaluación hasta ahora desarrollados.

A manera de conclusión

Se ha podido apreciar la complejidad en torno a la elaboración y administración de pruebas criteriales; no obstante, en la actualidad existe una gran cantidad de recursos conceptuales y metodológicos desarrollados para establecer líneas de investigación orientadas a la mejora continua y uso adecuado de los resultados de este tipo de pruebas educativas en la evaluación de competencias. Hasta ahora el trabajo de investigación en este campo es aún incipiente y desafortunadamente en nuestro país ni siquiera se considera importante, a pesar del uso creciente de pruebas criteriales en procesos de certificación de competencias. Los pocos trabajos de investigación que existen no se han difundido lo suficiente, a pesar de que de ellos depende en gran medida ganar credibilidad y con ello utilidad de los resultados generados durante casi una década de aplicación de pruebas internacionales y nacionales en nuestro país.

En la última década se ha evaluado a miles de egresados de las principales universidades públicas y privadas del país en diversos campos profesionales, lo que ha generado grandes bases de datos; pero es evidente que el uso que se ha hecho de los resultados es muy pobre y en ocasiones inapropiado, además de los problemas ocasionados por una difusión distorsionada, que lejos de apoyar los distintos niveles de decisiones produce errores conceptuales con implicaciones negativas para el propósito esencial de mejorar la calidad de la educación en nuestro país. En educación básica el panorama no es más alentador: la aplicación poblacional de la prueba ENLACE ha producido fuentes de invalidez por la falta de controles en su aplicación, afectando la interpretación y credibilidad de los resultados. En el INEE se han realizado esfuerzos importantes en el sentido planteado en este artículo, los cuales se han difundido en diversas publicaciones técnicas, no obstante éstas son consultadas por muy pocas personas.

Se requiere un mayor esfuerzo, dadas las implicaciones que actualmente tiene la evaluación a gran escala en México; se necesita invertir en la profesionalización del evaluador educativo y en el desarrollo de las líneas de investigación comentadas en este artículo, así como en mecanismos de difusión más eficaces para orientar a estudiantes, profesores, directivos y padres de familia en el uso adecuado de los resultados de estas pruebas para que realmente tenga sentido seguir aplicándolas a nivel nacional. Finalmente, y no menos importante, es imperativo que quienes son responsables de tomar decisiones para reorientar políticas públicas en materia de educación no sean tan ajenos a estos temas.

Referencias

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

Anderson, R.C. (1972), "How to Construct Achievement Tests to Assess Comprehension?", *Review of Educational Research*, núm 42, pp. 145-170.

Angoff, W.H. (1988), "Validity: An evolving concept", en H. Wainer y H.I. Braun (eds.), *Test Validity*, New Jersey, LEA, pp. 19-32.

Baxter, G.P. y R.J. Shavelson (1994), "Science Performance Assessments: Benchmarks and surrogates", *International Journal of Educational Research,* vol. 21, núm. 3, pp. 279-298.

Beaton, A.E. y N.L. Allen (1992), "Interpreting Scales Through Scale Anchoring", *Journal of Educational Statistics*, vol. 17, pp. 191-204.

Berk, R.A. (1976), "Determination of Optimal Cutting Scores in Criterion-referenced Measurement, *Journal of Experimental Education*, vol. 45, pp. 4-9.

Berk, R.A. (1978), "The Application of Structural Facet Theory to Achievement Test Construction", *Educational Research Quarterly*, vol. 3, pp. 62-72.

Berk, R.A. (1996), "Standard Setting: The next generation (where few psychometricians have gone before)", *Applied Measurement in Education*, vol. 9, núm. 3, pp. 215-235.

Beuck, C.H. (1984), "A Method for Reaching a Compromise between Absolute and Relative Standards in Examinations", *Journal of Educational Measurement*, vol. 21, pp. 147-152.

Biggs, J.B. y K.E. Collins (1982), *Evaluating the Quality of Learning: The SOLO Taxonomy,* Nueva York, Academic Press.

Brennan, R.L. y R.E. Cockwood (1980), "A Comparison of the Nedelsky and Angoff Cutting Score Procedures Using Generalizability Theory", *Applied Psychological Measurement*, vol. 4, núm. 2, pp. 219-240.

Cizek, G.J. (2005), "Adapting Testing Technology to Serve Accountability Aims: The case of vertically-moderated standard seeting", *Applied Measurement in Education*, vol. 18, núm. 1, pp. 1-10.

Cizek, G.J. y M.B. Bunch (2007), *Standard Setting: A guide to establishing and evaluating performance standards on tests*, California, SAGE Publications.

Cronbach, L.J. (1971), "Test Validation", en R.L Thorndike (ed.), *Educational Measurement*, Washington, American Council on Education, pp. 443-507.

Ebel, R.L. (1962), "Content Standard Test Scores", *Educational and Psychological Measurement*, vol. 22, pp. 15-25.

Embreston, S.E. y S.P. Reise (2000), *Item Response Theory for Psychologists*, New Jersey, LEA.

Glaser, R. (1963), "Instructional Technology and the Measurement of Learning Outcomes", *American Psychologist*, vol. 18, pp. 515-521.

Glaser, R. y A.J. Nitko (1971), "Measurement in Learning and Instruction", en R. Thorndike (ed.), *Educational Measurement*, Washington, American Council on Education, pp. 1040-1044.

Greeno, J.G. (1980), "Psychology of Learning, 1960-1980: One participant observation", *American Psychologist*, vol. 35, pp. 713-728.

Haertel, E. (1985), "Construct Validity and Criterion-Referenced Testing", *Review of Educational Research*, vol. 55, pp. 23-46.

Hambleton, R.K. (1980), "Test Score Validity and Standard-setting Methods", en R.A. Berk, *Criterion-Referenced Measurement: The state of the art*, Baltimore, Johns Hopkins University Press, pp. 80-123.

Hambleton, R.K. (1984), "Validating the Test Scores", en R. Berk (ed.), *A Guide to Criterion-Referenced Test Construction*, Baltimore, MD, The Johns Hopkins University Press, pp. 199-230.

Hambleton, R.K. (1985), "Criterion-Referenced Assessment of Individual Differences", en C.R. Reynolds y V.L. Willson (eds.), *Methodological and Statistical Advances in the Study of Individual Differences*, Nueva York, Plenum Press, pp. 393-424.

Hambleton, R.K. (1995), "Criterion-Referenced Measurement", en T. Husan y T.N. Postlethwaite (eds.), *International Encyclopedia of Education*, Nueva York, Pergamon Press, pp. 1182-1189.

Hambleton, R.K. y H. Swaminathan (1978), "Criterion-Referenced Testing and Measurement: A review of technical issues and developments", *Review of Educational Research,* vol. 40, pp. 1-47.

Hively, W., H.L. Patterson y S.A. Page (1968), "Universe-defined System of Arithmetic Achievement Tests", *Journal of Educational Measurement*, vol. 5, pp. 275-290.

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*, San Francisco, Jossey-Bass, pp. 109-127.

Jornet, J.M. y E. Backhoff (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, INEE, Colección Cuadernos de Investigación, núm. 30.

Jornet, J.M. y J. González (2009), "Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo", *Estudios sobre Educación*, núm. 16, pp. 103-123.

Jornet, J.M., J. González y J.M. Suárez (2010), "Validación de los procesos de determinación de estándares de interpretación para pruebas de rendimiento educativo", *Estudios sobre Educación* (en prensa).

Jornet, J.M. y J.M. Suárez (1994), "Evaluación referida al criterio: construcción de un *test* criterial de clase", en V. García Hoz, *Problemas y métodos de investigación en educación personalizada,* Madrid, Rialp, pp. 419-443.

Jornet, J.M. y J.M. Suárez (1987), "Un procedimiento para la determinación de estándares y establecimiento de puntos de corte en programas educativos", *Estudios de la Revista BORDON*, vol. 41, pp. 217-236.

Kane, M.T. (1982), "A Sampling Model of Validity", *Applied Psychological Measurement*, vol. 6, pp. 126-160.

Kieffer, K.M (1999), "Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate", *Advances in Social Science Methodology*, vol. 5, pp. 149-170.

Kingstone, N.M., S.R. Kahl, K. Sweeney y L. Bay (2001), *Setting Performance Standards: Concepts, methods and perspectives*, Mahwah, N.J., Lawrence Erlbaum.

Leyva, Y.E. (2004), *Validez de constructo en la evaluación de competencias médicas mediante pruebas referidas a criterio*, Tesis doctoral, México, Universidad Autónoma de Aguascalientes.

Lewis, D.M., H.C. Mitzel y D.R. Green (1996), "Standard Setting: A bookmark approach", en D.R. Green (dir.), *IRT-Based Standard-Setting Procedures Utilizing Behavioural Anchoring*, simposio organizado por el Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, junio.

Linn, R.L. (1979), "Issues of Validity in Measurement for Competency-Based Programs", en M.A. Bunda y J.R. Sanders (eds.), *Practices and Problems in Competency-Based Measurement*, Washington, National Council on Measurement in Education, pp. 547-561.

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

Marzano, R.J. (2007), *Designing a New Taxonomy of Educational Objectives*, California, Corwin Press, Inc., Sage Publications Company.

Messik, S. (1975), "The Standard Problem: Meaning and values in measurement and evaluation", *American Psychologist*, vol. 30, pp. 955-966.

Millman, J. (1974), "Criterion-Referenced Measurement", en W.J. Popham (ed.), *Evaluation in Education: Current applications*, Berkeley, McCutchan, pp. 205-216.

Patel, V.L., D.R. Kaufman y J.F. Arocha (2000), "Conceptual Change in the Biomedical and Health Sciences Domain", en R. Glaser (ed.), *Advances in Instructional Psychology,* vol. 5: *Educational Design and Cognitive Science*, Londres, LAE, pp. 329-392.

Plake, B.S., R.K. Hambleton y R.M. Jaeger (1997), "A New Standard Setting Method form Performance and Assessment. The dominant profile judgment method and some field-test results", *Educational and Psychological Measurement*, vol. 57, pp. 400-411.

Popham, W.J. (1983), *Evaluación basada en criterios*, Madrid, Magisterio Español, S.A.

Popham, W.J. y T.R. Husek (1969), "Implication of Criterion-Referenced Test", *Applied Psychology Measurement*, vol. 4, pp. 469-492.

Ruiz-Primo, M.A. y R.J. Shavelson (2001), "Comparison of the Reliability and Validity of Scores from two Concept-Mapping Techniques", *Journal of Research in Science Teaching*, vol. 38, núm.2, pp. 260-278.

Scandura, J.M. (1977), *Problem Solving: A structural/process approach with educational implications*, Nueva York, Academic Press.

Schoon, C.G., C.M. Guillion y P. Ferrara (1979), "Bayesian Statistics, Credentialing Examinations and the Determination of Passing Points", *Evaluation and the Health Professions*, vol. 2, pp. 181-201.

Shavelson, R.J. y M.A. Ruiz-Primo (2000), "On the Psychometrics of Assessing Science Understanding", en J. Mintzes, J. Wandersee y J. Novak (eds.), *Assessing Science Understanding*, San Diego, Academic Press, pp. 303-341.

Shepard, L.A. (1980), "Standard Setting Issues and Methods", *Applied Psychological Measurement*, vol. 4, núm. 4, pp. 447-467.

Sireci, S.G., R.K. Hambleton y M.J. Pitoniak (2004), "Setting Passing Scores on Licensure Examinations Using Direct Consensus", *CLEAR Exam Review*, vol. 15, núm. 1, pp. 21.25.

Snow, R.E. (1980), "Aptitude and Achievement", en W.B. Schrader (ed.), *Measuring Achievement: Progress over a decade, new directions for testing and measurement*, San Francisco, Jossey Bass, pp. 47-103.

Van der Linden, W.J. (1984), "Decision Models for the Use with Criterion-Referenced Tests", *Applied Psychological Measurement*, vol. 4, pp. 469-492.

Zeller, R.A. (1988), "Validity", en J.P. Keeves (ed.), *Educational Research, Methodology and Measurement: An International Handbook*, Nueva York, Pergamon Press, pp. 322-330.