PARADIGMA, Vol. X, No. 1 — JUNIO DE 1989

VALIDEZ, CONFIABILIDAD Y ESTABLECIMIENTO DE NIVELES DE RENDIMIENTO EN PRUEBAS BASADAS EN DOMINIOS

Hernando Salcedo Galvis
Escuela de Educación
Facultad de Humanidades y Educación
Universidad Central de Venezuela

RESUMEN

El propósito de este trabajo fue analizar algunos de los avances recientes en relación con tres áreas de interés en el campo de las pruebas basadas en criterios (PBC), denominadas por algunos autores pruebas basadas en dominios (PBD): la validez, la confiabilidad y el establecimiento de niveles de rendimiento o estándares. En el artículo se analizan algunas contribuciones fundamentales en relación con los temas tratados, especialmente lo relativo a los procesos de validación de una prueba basada en dominios, al concepto de variabilidad y sus implicaciones en la confiabilidad de PBD, y al concepto de consistencia de la decisión; se analizan igualmente algunos métodos para el establecimiento de estándares, y se concluye con algunas consideraciones en torno a la factibilidad de aplicación de las PBD en la educación venezolana.

Las pruebas basadas en criterios (PBC) o pruebas basadas en dominios (PBD), como se les denomina indistintamente, constituyen una tecnología susceptible de utilizar exitosamente en la educación venezolana. Estas pruebas han sido definidas por Popham (1978b) como un tipo especial de pruebas de rendimiento cuyo propósito es determinar la situación de un individuo con respecto a un dominio de conductas "bien definido", es decir, especificado de tal manera que proporcione una descripción clara del significado de la actuación que se evalúa.

Las pruebas basadas en dominios están asociadas al concepto de ***continuo*** de ejecución respecto de un dominio de conductas de grados diferentes de complejidad. Dentro de estas pruebas, existe una clase particular, las pruebas de competencias ("mastery"), cuyo propósito es clasificar individuos en términos de establecer si poseen o no ciertas competencias básicas, en función de un puntaje de corte o estándar.

Según Hambleton, Swaminathan, Algina y Coulson (1978), y Subkoviak y Baker (1977), existen dos propósitos de las pruebas basadas en criterios: estimar la proporción de puntajes correctos de un individuo, y tomar decisiones respecto de la presencia o no en un individuo de ciertas competencias ("mastery" – "nonmastery"), en cuyo caso se requiere un puntaje de corte. Una clase particular de pruebas de competencias son las denominadas ***pruebas de competencias mínimas***, cuyo propósito es clasificar individuos en dos categorías: ***aprobados*** o ***no aprobados***, pero en las cuales el concepto de competencias mínimas es clave, ya que aunque no difiere esencialmente de la noción de competencias en general, se refiere a competencias "consideradas absolutamente necesarias" (Shepard, 1980, p. 34) respecto de un curso de acción inmediato, por ejemplo, el ingreso de un individuo a la universidad, o al ejercicio de una profesión, y distingue competencias que son esenciales de aquéllas que no lo son.

La tecnología de las pruebas basadas en dominios ha avanzado en los últimos veinte años hasta llegar a un grado de consistencia tal que permite utilizar estos instrumentos para mejorar la práctica educativa,

aunque persisten aun dificultades técnicas y conceptuales no resueltas satisfactoriamente. En consecuencia, el propósito de este trabajo es analizar algunos avances recientes relacionados con tres de los mayores problemas confrontados por las pruebas basadas en dominios: la ***validez***, la ***confiabilidad*** y el establecimiento o fijación de ***niveles de rendimiento o "estándares"***.

VALIDEZ

El concepto de validez es de importancia crucial en relación con un instrumento de medición. Sin embargo, las nociones convencionales de validez, asociadas con la "teoría clásica de los test", resultan insuficientes y requieren ser revisadas cuando se trata de las pruebas basadas en criterios o dominios, y más específicamente, de pruebas de competencias mínimas. Estas insuficiencias han sido estudiadas por numerosos autores, por ejemplo, Cronbach (1971), Cronbach, Gleser, Nanda y Rajaratnam (1972), Hambleton (1982), Hambleton et al. (1978), Linn (1979, 1980), Popham (1975, 1978b, 1981), Popham y Husek (1969).

Según Shepard (1980b), en el proceso de validación de una prueba de competencias mínimas, debe atenderse a las diferentes etapas de su construcción, es decir, la ***selección*** del dominio a evaluar, su ***definición y especificación***, y la ***selección de los items o preguntas*** que se supone representan el dominio seleccionado. En relación con la ***selección*** del dominio, ésta debe iniciarse antes de prescribir especificaciones para la prueba, y no como suele hacerse (ver al respecto, Hambleton et al., 1978; Millman, 1974; Popham, 1975, 1978b), bajo el supuesto de que el dominio ha sido seleccionado previamente en forma adecuada. La atención que se le atribuya a la selección del dominio como previa a su especificación, adquiere mayor importancia respecto de objetivos a largo plazo, en cuyo caso no existe certeza acerca de qué competencias deberían ser evaluadas. Además, todos los tipos tradicionales de validez son requeridos en las pruebas basadas en dominios. Así, la ***validez de contenido*** tiene que ver con el grado en que los items de la prueba

constituyen una muestra representativa del dominio de conductas de interés. Respecto de la ***validez predictiva***, según Subkoviak y Baker (1977), "cuando una prueba es diseñada con el propósito de clasificar individuos como poseedores o no poseedores de ciertas competencias ("masters or nonmasters"), existe generalmente la creencia de que los resultados de la prueba están relacionados con resultados en otras variables tales como éxito o fracaso en tareas subsecuentes. En este caso, la validez predictiva, o el grado de asociación entre los resultados de la prueba y otras variables criterio, constituye una consideración básica" (p. 294).

En cuanto a la ***validez de constructo***, ésta es necesaria cuando se trata de hacer inferencias basadas en los resultados obtenidos en una prueba, "acerca de una capacidad no manifiesta, es decir, cuando las conductas representadas en la prueba no son directamente de interés sino sólo representaciones o indicios respecto de los comportamientos que sirven como criterio. En este caso, la relación teórica entre la ejecución en la prueba y otros comportamientos, debe ser confirmada por la realidad" (Shepard, 1980b, p. 39).

Como se dijo en otra oportunidad respecto de la validez de constructo (Salcedo, 1986), si se interpretaran los puntajes de una prueba como evidencia de un constructo tal como comprensión lectora o habilidad numérica, esta interpretación implicaría que existe un rasgo o atributo el cual puede ser denominado ***comprensión lectora*** o ***habilidad numérica***, al que puede atribuirse el desempeño del individuo en la prueba, al menos hasta cierto grado.

En cuanto a la ***definición*** del dominio, las especificaciones del dominio a ser evaluado aseguran la inclusión en la prueba de conductas consideradas básicas, lo cual permite juzgar su contribución lógica a la validez de ***constructo***, y además, la congruencia entre los ítems de la prueba y el dominio. Al respecto, numerosos autores han propuesto procedimientos para la especificación de dominios, por ejemplo, Hively y asociados (1973), Millman (1974) y Popham (1975, 1978b, 1981). El procedimiento propuesto por Popham, descrito en detalles

en trabajos anteriores (Salcedo, 1980, 1986), es uno de los menos complejos, quizá el más difundido, y ha sido utilizado en Venezuela por el autor en sus cursos de evaluación a nivel de postgrado.

Otro procedimiento para la elaboración y validación de pruebas basadas en dominios es el propuesto por Burry, Herman y Baker (1984), tratado previamente por Baker y Herman (1983). Este procedimiento se basa en resultados obtenidos en investigaciones realizadas en el Centro para el Estudio de la Evaluación de la Universidad de California (UCLA CSE), y supera numerosas limitaciones de métodos propuestos previamente. El nuevo enfoque incluye seis componentes en las ***especificaciones*** de un dominio, los cuales se describen brevemente a continuación.

1. ***Descripción del dominio***, la cual se refiere a la ejecución que se espera de un sujeto en un área determinada, y puede consistir en un objetivo o en la descripción de una tarea y sus componentes.

2. ***Límites del contenido***, los cuales establecen el rango de contenido que puede ser usado para elaborar los ítems o preguntas de la prueba, es decir, el ***contenido elegible***. Este componente varía según se trate de pruebas para ***seleccionar*** las respuestas (las denominadas "pruebas objetivas"), o de pruebas para ***elaborar*** las respuestas (pruebas de ensayo).

3. ***Límites de los distractores o criterios de respuesta***. Describen las alternativas ***incorrectas*** que pueden ser usadas para respuestas de selección, o los criterios que proporcionan las reglas mediante las cuales se juzgan las respuestas elaboradas por el estudiante en pruebas de ensayo.

4. ***Formato***, el cual se refiere a la descripción de la forma en que los ítems o preguntas pueden ser presentados.

5. ***Instrucciones.*** Este componente se refiere al conjunto específico de reglas que orientan al estudiante para responder las preguntas o ítems de una prueba.

6. ***Ítem de muestra.*** Responde a las reglas establecidas en los cinco componentes anteriores, y constituye una guía a seguir por quienes elaboran los ítems de una prueba.

La validez de las pruebas elaboradas según el procedimiento descrito previamente depende en gran medida del grado en que las especificaciones de un dominio de contenido se correspondan con los procesos de instrucción y evaluación.

A fin de hacer más rigurosa y precisa la correspondencia entre las especificaciones de una prueba y los ítems o preguntas que la integran, este enfoque es complementado con una escala para evaluar la congruencia de cada ítem con el dominio o estructura a la cual corresponde. Esta escala proporciona un rango de valores que permiten juzgar con mayor precisión el grado de correspondencia ítem–dominio en función de cada una de las categorías descritas en las especificaciones de la prueba.

Además de proporcionar esta escala, la cual supera las limitaciones implícitas en el uso de escalas dicotómicas (congruencia–incongruencia), el enfoque propuesto por Burry y asociados (1984) contempla en dicha escala dos categorías adicionales las cuales contribuyen a afinar aun más el proceso de validación. Estas categorías se refieren a ***complejidad lingüística*** y ***complejidad del pensamiento.*** La complejidad lingüística tiene que ver con el significado del vocabulario usado en los ítems, así como con la estructura del lenguaje y su correspondencia con las especificaciones de la prueba. La complejidad del pensamiento se refiere a lo siguiente: (a) aquellos procesos mentales requeridos para responder la pregunta, pero que no son incluidos en la descripción del dominio o en los límites del contenido (por ejemplo, legibilidad de la escritura, capacidad de memoria a corto plazo, imaginación, etc.), y que se supone están al alcance de todos los estudiantes a cierto nivel

de competencia; (b) las instrucciones para responder la pregunta proporciona la misma cantidad de información y estructura para todos los estudiantes; y (c) en el caso de preguntas con componentes no verbales, se considera razonable suponer que tales componentes se corresponden con los límites del contenido o los límites de los distractores en su significado, y que esta interpretación es estable a través de todos los grupos de individuos que toman la prueba.

Respecto de la ***selección y análisis de ítems,*** otra etapa clave en la construcción de las pruebas basadas en dominios, la validez de contenido de una prueba depende de la correspondencia entre las ***especificaciones*** del dominio a evaluar y los ítems o preguntas que representan dicho dominio. En general, existen dos enfoques en el análisis de ítems en este tipo de pruebas: procedimientos ***a priori,*** basados en juicios de expertos, y procedimientos a ***posteriori,*** o empíricos. En el primer caso, se trata de determinar el grado de ***congruencia*** existente entre los ítems de la prueba y las especificaciones del dominio, lo cual se realiza mediante la utilización de expertos en contenido, quienes deben establecer en qué grado el formato y contenido de un ítem reflejan adecuadamente un aspecto particular considerado en las especificaciones del dominio (Hambleton, 1980; Popham, 1975, 1978b).

El segundo enfoque consiste en aplicar técnicas empíricas o estadísticas en forma parecida a las empleadas en pruebas basadas en normas. En general, estos procedimientos deben ser aplicados e interpretados con precaución, dadas las limitaciones y problemas que implican cuando se utilizan en el contexto de las pruebas basadas en dominios. Sin embargo, cuando el propósito de la prueba es determinar grado de competencia, es decir, discriminar individuos en poseedores y no poseedores de ciertas competencias, en cuyo caso la prueba debe tener validez de constructo, los ítems deben ser seleccionados de manera que discriminen con exactitud entre las dos categorías de individuos. Con este propósito, se han desarrollado numerosos índices de discriminación, entre los cuales quizá el más conocido es el "índice de sensitividad instruccional" (ISI), definido por Haladyna y Roid (1983,p. 36) como "la tendencia de los ítemes a variar en dificultad como una fun-

ción de la instrucción". Aunque existen varios enfoques para el estudio del ISI, y los procedimientos estadísticos son complejos, la investigación ha demostrado que el ***índice de diferencia pretest-postest*** (IDPP) es el más simple conceptual y empíricamente, así como fácil de calcular, por lo que según estos autores es el más recomendable para la revisión empírica de ítems en pruebas basadas en criterios.

CONFIABILIDAD

La ***confiabilidad*** en las pruebas basadas en criterios es otro de los tópicos que ha sido objeto de atención durante las dos últimas décadas. Ejemplos de tal atención son los trabajos de Popham y Husek (1969), Swaminathan, Hambleton y Algina (1974), Hambleton y asociados (1978), y Subkoviak (1982), entre otros, de consulta obligada en un análisis de este problema. A continuación se presenta una síntesis de los avances más importantes.

En su célebre artículo publicado en 1969, Popham y Husek establecieron que los métodos para determinar la confiabilidad, basados en la teoría clásica de los tests, eran inapropiados en el caso de las pruebas basadas en criterios, debido a que tales métodos dependen de la ***variabilidad*** de los puntajes, mientras que en las pruebas basadas en criterios la variabilidad es irrelevante. En 1978 Hambleton y asociados sugirieron una solución al problema de la falta de varianza, mediante la utilización de grupos de individuos considerados poseedores y no poseedores ("competentes" e "incompetentes") ***respecto del material incluido en una prueba***. Por su parte, Subkoviak y Baker (1977) destacaron los dos propósitos básicos de las PBC respecto de la confiabilidad: determinar la ***consistencia de las estimaciones en términos de la proporción correcta de puntajes de un individuo*** en varias aplicaciones de la misma prueba o de pruebas equivalentes, o ***clasificar individuos*** en grupos de "competentes" e "incompetentes" ("masters" y Nonmasters"), en cuyo caso la confiabilidad "se refiere a la consistencia o estabilidad de tales clasificaciones en pruebas repetidas" (p. 287).

En relación con el primer propósito, Harris (1972) y Cronbach y asociados (1972) han propuesto alternativas a la falta de variabilidad tales como el uso del error estándar de medida y la teoría de la "generalizabilidad", respectivamente. Respecto del segundo propósito, Hambleton y Novick (1973) introdujeron el concepto de "consistencia de la decisión" y "sugirieron usar la proporción observada de acuerdo como un índice de confiabilidad" (Shepard, 1980b, p. 47).

A partir del concepto de ***consistencia de la decisión***, Swaminathan y asociados (1974) propusieron que la proporción de individuos clasificados consistentemente como "master/master" o "nonmaster/nonmaster" en dos aplicaciones de una prueba, sea considerada como un índice grupal de confiabilidad. Estos autores introdujeron una corrección respecto de la proporción observada de acuerdo (Po) que ocurriría por azar (pa), para lo cual adoptaron el coeficiente Kappa (k) de Cohen como un índice de confiabilidad ajustado respecto del azar. De manera que la fórmula a utilizar en el cálculo de la confiabilidad será la siguiente:

$$k = \frac{Po - Pa}{1 - Pa}$$; en la cual:

Po es la proporción de acuerdo, Pa es la proporción de acuerdo que ocurriría sólo por azar, y k es la proporción de clasificaciones consistentes más allá de la proporción de acuerdo por azar atribuible a las proporciones particulares de individuos considerados como competentes e incompetentes en las pruebas respectivas (Subkoviak, 1982).

Además del método propuesto por Swaminathan y asociados, el cual requiere dos aplicaciones de la misma prueba o de formas equivalentes, se han propuesto otros métodos para determinar la ***consistencia de la decisión*** cuando se clasifican individuos en "competentes" e "incompetentes". Entre los más conocidos están el de Carver (1970), el de Huynh (1976), el de Subkoviak (1976), y el de Marshall y Haertel (1976). El método de Carver requiere, como el de Swaminathan y

asociados, la aplicación repetida de la misma prueba o de formas equivalentes, y no es recomendable debido a que no es sensible a la consistencia de las clasificaciones individuales, es decir, en cada aplicación, el grupo de individuos clasificados como "competentes" puede ser diferente (Subkoviak, 1982).

En cuanto a los métodos propuestos por Huynh, Subkoviak, y Marshall y Haertel, éstos requieren la aplicación sólo de una prueba, lo cual los hace recomendables en comparación con el método de Swaminathan y asociados. En un estudio comparativo realizado por Subkoviak (1978), éste concluyó que estos métodos "pueden suministrar estimaciones razonablemente precisas de la proporción de clasificaciones consistentes en dos pruebas de competencia" (p. 115). Sin embargo, cada uno de estos métodos presenta ventajas y desventajas. Así, el método de Swaminathan y asociados es computacionalmente simple y además produce estimadores insesgados, pero requiere dos pruebas o mediciones, y los errores estándar tienden a ser relativamente grandes para muestras cercanas a 30. Respecto de los métodos de Huynh, Subkoviak, y Marshall y Haertel, éstos tienen las ventajas de que requieren sólo la aplicación de una prueba y los errores estándar de estimación son relativamente pequeños, pero son tediosos computacionalmente y producen estimaciones sesgadas para pruebas cortas. Subkoviak concluyó recomendando el método de Huynh, ya que éste "requiere sólo una prueba y produce estimaciones razonablemente precisas, las cuales parecen ser ligeramente conservadoras para pruebas cortas" (p. 115).

EL ESTABLECIMIENTO DE NIVELES DE RENDIMIENTO O ESTANDARES

El tercer aspecto investigado en este estudio se refiere al establecimiento de niveles de rendimiento, ***estándares o puntajes de corte*** en pruebas basadas en dominios. Este constituye quizá el problema más

controversial dentro de este campo, como lo evidencian las numerosas publicaciones acerca del tema durante las dos últimas décadas. El carácter controversial del problema obedece a sus múltiples implicaciones de índole técnica, social, ética, y política. (Al respecto, ver por ejemplo, los puntos de vista de Block, 1978; Burton, 1978; Glass, 1978; Hambleton, 1978; Linn, 1978; Popham, 1978a; Scriven, 1978; Shepard, 1980a).

Concepto y Tipos de Estándares

Según Hambleton (1982), "un estándar es un punto en una escala de puntajes el cual es usado para clasificar examinandos en dos categorías que reflejan niveles diferentes de competencia en relación con un objetivo particular (o conjunto de objetivos) medido por una prueba" (pp. 99–100). Existen tres tipos de estándares o niveles de competencia asociados con el uso de las pruebas basadas en criterios, los cuales son expresados en términos de porcentajes. El primer tipo se refiere al nivel de logro del estudiante respecto de cada uno de los objetivos o ***dominios*** evaluados por una prueba; el segundo tipo se refiere al porcentaje de objetivos o dominios que debe lograr el estudiante respecto del total de objetivos evaluados por la prueba; y finalmente, el tercer tipo de estándar se refiere al porcentaje de estudiantes de un nivel determinado que deben lograr los objetivos evaluados por la prueba. (Ver al respecto Hambleton, 1978, 1982; Popham, 1975).

Métodos para el Establecimiento de Estándares

Existen métodos diversos para el establecimiento de estándares o niveles de rendimiento. Hambleton (1982) y Hambleton y Eigor (1980) distinguen tres categorías: métodos basados en juicios, métodos combinados y métodos empíricos. Pero independientemente del grado de complejidad y sofisticación de estos métodos, todos ellos son arbitrarios en mayor o menor grado, ya que suponen la presencia de juicios de una clase u otra. Al respecto, Popham (1981) ha establecido

una distinción muy útil entre lo que es arbitrario en el sentido positivo, es decir, basado en la decisión de un juez calificado para ello, y lo que denota actuar caprichosamente, sin razón ni criterio definido. Según Popham, la arbitrariedad en el sentido positivo es lo que debe caracterizar el establecimiento de niveles de rendimiento.

Con el propósito de analizar diversos métodos para establecer estándares, Meskauskas (1976) distinguió dos categorías de modelos: los que conciben el nivel de ejecución como un punto de un "continuo" ("continuum models"), y los que conciben dicho nivel como "todo o nada" ("state models"). Son características de los primeros las siguientes: (a) el dominio ("mastery") significa una habilidad o conjunto de habilidades distribuidas en forma continua; (b) un área es identificada en el extremo superior del continuo, y si un individuo iguala o excede el límite inferior de ésta, se considera que ha logrado el nivel de ejecución deseado; y (c) el propósito del proceso de medida es obtener información para la toma de decisiones en términos dicotómicos.

Respecto de los modelos basados en la noción de "mastery" como una expectativa de ***todo –o –nada***, sus características son: (a) la ejecución en términos de puntajes verdaderos es una tarea dicotómica de todo–o–nada; (b) el estándar en términos de puntajes verdaderos es fijado en el ***cien por ciento***; y (c) después de considerar los errores de medida, generalmente se adoptan estándares inferiores al cien por ciento. Aunque los modelos de ***todo o nada*** son aplicables en ciertos casos relacionados con destrezas físicas y cognitivas, para la mayoría de destrezas cognitivas son inapropiados debido al carácter continuo de éstas.

En pruebas cuyo propósito es clasificar individuos en dos categorías ("competentes" e "incompetentes"), caso en el cual se requiere establecer un nivel de rendimiento o estándar, sólo son aplicables los modelos basados en el concepto de ***continuo***. Esto significa que la habilidad distribuida a lo largo de ese continuo debe ser dicotomizada arbitrariamente, pero en el sentido positivo del término "arbitrario" propuesto por Popham (1981), es decir, reflexiva y sistemáticamente.

A continuación se presenta una síntesis de algunos de los métodos más conocidos para el establecimiento de niveles de rendimiento o estándares.

Métodos Basados en Juicios

En los métodos basados en juicios, los ítems individuales son examinados por varios jueces a fin de determinar cuál sería la actuación en cada ítem de la prueba de una persona mínimamente competente.

El Método de Nedelsky (1954). Este método, quizá el más antiguo, se usa con pruebas de ***opción múltiple***, y se basa en el análisis de los distractores o alternativas incorrectas. La aplicación de este método incluye las etapas siguientes: (a) designar jueces familiarizados con las competencias que se desea evaluar y su dominio por los estudiantes; (b) consideración por los jueces de los distractores para cada ítem, e identificación de aquellos distractores reconocibles como incorrectos por un estudiante ***mínimamente competente***; (c) conversión, para cada ítem, de las respuestas ***no eliminadas en la segunda etapa***, es decir, la respuesta correcta más los distractores no eliminados, a una probabilidad de "correcto por azar" (por ejemplo, 2 respuestas no eliminadas correspondería a una probabilidad de 0.50, 3 respuestas a una probabilidad de 0.33, etc.); y (d) sumar estas probabilidades de "correcto – por – azar" para cada juez y luego promediarlas para obtener un ***estándar*** de ejecución para los estudiantes ***mínimamente competentes***.

El Método de Angoff (1971). Este método, conocido como la ***adaptación de Angoff***, es ligeramente más fácil que el método de Nedelsky, y además puede ser usado con cualquier tipo de ítems y no sólo ítems de opción múltiple. Aunque esencialmente similar al método de Nedelsky, en este caso los jueces consideran cada ítem y estiman la probabilidad de que un estudiante ***mínimamente competente*** lo responda correctamente. Estas probabilidades son luego sumadas y promediadas, obteniéndose así el puntaje de aprobación o estándar. En caso de usar este método con pruebas de opción múltiple, un juez no debería nunca asignar a un ítem una probabilidad menor que la que

resultaría por azar según el número de opciones. Así, en una prueba con ítems de cuatro opciones, la probabilidad asignada nunca debería ser inferior a 0.25, es decir, una de cuatro.

El Método de Ebel (1972). Este método es similar al método de Angoff, con la diferencia de que los jueces deben categorizar los ítems según su ***relevancia*** y ***dificultad***, usando cuatro niveles de relevancia (esencial, importante, aceptable y cuestionable) y tres niveles de dificultad (fácil, medio y difícil). Estos niveles forman así una tabla de 3 x 4. Las probabilidades de responder correctamente un ítem son asignadas a cada categoría y usadas para ponderar los ítems en la determinación del puntaje mínimo de aprobación. Específicamente, lo que hacen los jueces es ubicar cada uno de los ítems en la celda correspondiente, en función de su relevancia y dificultad, y asignar un porcentaje a cada celda (el porcentaje de ítems que el examinando mínimamente calificado sería capaz de responder). Luego, el número de ítems en cada celda es multiplicado por el porcentaje acordado por los jueces. La suma de todas las celdas dividida entre el número total de ítems representa el estándar.

Métodos Combinados

En los métodos combinados (combinación de datos basados en juicios y datos empíricos), se pide a los jueces que emitan juicios acerca de los niveles de experticia ("mastery") de los estudiantes, en lugar de juicios acerca de los ítems de una prueba.

El Método del Grupo Limítrofe. Este método, propuesto por Zieky y Livingston (1977), requiere que los jueces identifiquen estudiantes cuya experticia respecto de la competencia en cuestión está tan cerca del límite entre aceptable e inaceptable, que no pueden ser clasificados en ninguno de los dos grupos. Luego, se aplica a estos estudiantes la prueba mediante la cual se ha de determinar su competencia, y el puntaje correspondiente a la ***mediana*** es tomado como el estándar.

La aplicación de este método incluye las etapas siguientes: (a) selección de jueces familiarizados con la población de estudiantes involucrados; (b) discusión por los jueces de lo que constituye un ***rendimiento mínimamente aceptable***; (c) identificación de estudiantes considerados limítrofes; (d) aplicación de la prueba; y (e) cálculo de la ***mediana*** del rendimiento.

Un aspecto fundamental en este método es la selección de jueces conocedores de los niveles reales de rendimiento de los estudiantes, quienes han de seleccionar a aquellos cuyo rendimiento es considerado ***limítrofe***. Igualmente, tales jueces deben ser conocedores de las competencias para las cuales se desea establecer estándares. En la aplicación de este método se requiere un número suficiente de estudiantes limítrofes, generalmente unos cien, a fin de asegurar la confiabilidad del estándar.

El Método de Grupos Contrastados. El método de grupos contrastados, también propuesto por Zieky y Livingston (1977), requiere como el método del grupo límitrofe, que los jueces tengan conocimiento previo acerca del rendimiento de los estudiantes a considerar. Así, una vez que los jueces han definido el nivel de rendimiento mínimo aceptable para el área o asignatura de que se trate, identifican aquellos estudiantes claramente ubicables como poseedores o no poseedores de las competencias deseadas, excluyendo los individuos limítrofes. Luego, se aplica la prueba a ambos grupos y sus resultados se presentan en un gráfico en el cual el punto de intersección de las curvas correspondientes a los dos grupos representa el estándar (Ver Figura 1).

Aunque no es necesario que los dos grupos sean iguales, el grupo menor debería tener 100 individuos, aproximadamente, a fin de asegurar cierta estabilidad en la estimación.

Cuando se usa este método, es posible ajustar el estándar o punto de intersección de las curvas correspondientes a los dos grupos contrastados, desplazándolo hacia arriba o hacia abajo, a fin de reducir el nú-

Figura 1. Punto de intersección de los grupos que han logrado y no han logrado las competencias.

mero de "falsos positivos" (estudiantes identificados en la prueba como poseedores de las competencias previstas en los objetivos pero quienes ***no han logrado*** tales objetivos) o "falsos negativos" (estudiantes identificados como que no han logrado los objetivos pero que ***sí los han logrado***). (Ver Figura 2). La dirección en que sea desplazado el estándar o punto de corte dependerá de la importancia que se atribuya a los dos tipos de error, en función del contexto en que se haya de tomar la decisión. Es decir, el problema consiste en considerar cuáles serían las consecuencias de ***aprobar*** erróneamente a un estudiante que no ha logrado ciertas competencias, o de ***desaprobarlo*** erróneamente habiendo logrado tales competencias.

El Método de Juicio Basado en Información. Este método, propuesto por Popham (1981), se basa en la consideración de un conjunto de factores claves en el establecimiento de niveles de competencia, entre los cuales destaca la importancia que tiene el ***contexto de la decisión.*** Las etapas a considerar en la aplicación de este método son las siguientes:

1. ***Análisis del contexto de la decisión.*** Es decir, consideración de la magnitud e importancia de las decisiones a tomar y la situación en la cual esas decisiones han de ser tomadas. Se trata aquí de considerar las consecuencias que se derivan si un individuo no logra el estándar propuesto, y las implicaciones de los "falsos positivos" y "falsos negativos", es decir, las posibilidades que existen de aprobar erróneamente individuos que no han logrado los objetivos propuestos, o desaprobar erróneamente sujetos que han logrado tales objetivos, respectivamente.

2. ***Especificación clara de las competencias respecto de las cuales se han de establecer estándares.*** Esto puede lograrse mediante el análisis de los instrumentos que se utilizan para evaluar dichas competencias, es decir, a través de la revisión de las ***especificaciones*** de los dominios incluidos en una prueba. En este caso, el juicio de expertos en relación con el grado de congruencia entre tales especificaciones y las competencias que se desea evaluar adquiere importancia fundamental.

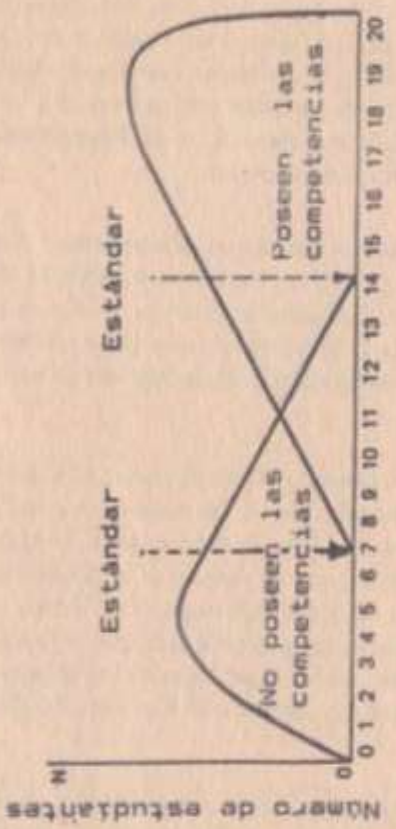

Figura 2. Estándares fijados para reducir los "falsos positivos" y "falsos negativos".

3. ***Obtención de información respecto del rendimiento de los siguientes tipos de individuos:*** (a) ***no expuestos aún al proceso de instrucción***, por ejemplo, estudiantes que están iniciando el séptimo grado de educación básica, a fin de obtener una idea general acerca de un límite inferior a partir del cual proyectar expectativas más realistas; (b) ***individuos sometidos recientemente a la instrucción***, por ejemplo, estudiantes de noveno grado de educación básica a quienes se les aplica una prueba de rendimiento en matemática durante el mes de mayo, y a partir de los resultados obtenidos según diferentes estándares, considerar la conveniencia de desplazar, hacia arriba o hacia abajo, el estándar deseado; y (c) individuos ***previamente expuestos al proceso de instrucción***, es decir, adultos de quienes se supone ejercitan las destrezas que se desea evaluar, lo cual puede ser de utilidad en el sentido de contrastar las expectativas de quienes establecen los estándares con las competencias requeridas en la vida real.

4. ***Obtención de información respecto de preferencias de grupos interesados.*** Este aspecto se refiere a la conveniencia de obtener información de grupos o personas cuyo interés puede ser de utilidad en el establecimiento de estándares. Tal sería el caso, por ejemplo, de padres y representantes, empleadores, expertos en curriculum e instrucción y otros expertos, dirigentes magisteriales y supervisores.

5. ***Establecimiento de estándares.*** Este es el último paso en el procedimiento propuesto, y se basa en la información obtenida mediante las etapas y procesos descritos antes.

Puede concluirse, en relación con los diversos métodos presentados, que la decisión respecto de cuál método emplear dependerá de las circunstancias y propósitos específicos para los cuales se establecen los estándares, así como de las ventajas y limitaciones inherentes a cada método. En todo caso, las decisiones que se tomen al establecer estándares serán siempre basadas en juicios, no obstante la presencia, en algunos métodos, de procedimientos matemáticos más o menos complejos.

Otro aspecto fundamental a considerar cuando se establecen estándares es el relativo a la decisión a tomar respecto de si se han de establecer niveles de rendimiento para ***cada competencia*** o si se fijará un sólo estándar que represente un porcentaje agregado respecto de ***varias competencias***. La decisión en este caso dependerá de las ventajas y desventajas presentes en una y otra alternativa. Así, el adoptar un puntaje agregado como estándar para varias competencias constituye un procedimiento más sencillo, y además el mayor número de ítems en la prueba total se manifestará en un grado mayor de confiabilidad que si se utilizaran pruebas más cortas por ***cada competencia***. Pero, por otra parte, el puntaje agregado para competencias diferentes presenta la desventaja de que no permite establecer el nivel de logro alcanzado por un individuo en cada competencia, lo cual despoja al estándar de su potencial poder diagnóstico respecto de áreas específicas, información que es de gran utilidad en relación con el mejoramiento del proceso de instrucción.

Cuando se establecen estándares por cada competencia, es necesario decidir acerca de si tales estándares serán iguales o diferentes, y si es necesario lograr ***todas*** las competencias o ***sólo una proporción*** de ellas. Es posible en ciertos casos fijar un estándar uniforme para varias competencias, pero no siempre esto es conveniente, dadas las diferencias que pueden existir en cuanto a su dificultad e importancia.

En cuanto a la decisión respecto de la proporción de competencias (u objetivos) a lograr, Popham (1981) sugiere que tal decisión sea tomada después de establecer estándares para cada competencia, ya que si se fijan niveles de exigencia altos para las competencias individuales, podría haber mayor flexibilidad en la decisión que si tales niveles son bajos.

EL USO DE PRUEBAS BASADAS EN DOMINIOS EN LA EDUCACION VENEZOLANA

El propósito de este estudio fue analizar algunos de los avances recientes en el campo de las pruebas basadas en dominios (PBD), a fin

de derivar conclusiones que orienten su uso en la educación venezolana. El análisis realizado permite formular las consideraciones siguientes:

1. La tecnología de las pruebas basadas en criterios o dominios ha alcanzado un grado de desarrollo el cual hace factible su utilización exitosa en Venezuela, si se toman en consideración las características, necesidades y limitaciones del sistema educativo en general y de algunas áreas y niveles en particular.

2. Innovaciones tales como la educación básica, la educación a distancia en sus diversas modalidades existentes en el país, y otras innovaciones en diferentes niveles del sistema educativo, requieren que los problemas de carácter conceptual, técnico e instrumental asociados con su implementación, sean tratados racionalmente, a fin de evitar la pérdida de esfuerzos y recursos que implican tales innovaciones. Uno de estos problemas se refiere a la manera como se realiza la evaluación del rendimiento estudiantil. La utilización apropiada de pruebas basadas en dominios puede contribuir a mejorar sustancialmente las prácticas evaluativas actuales, caracterizadas por su inconsistencia, la falta de criterios bien fundamentados teórica y técnicamente, y la ausencia de lineamientos claros respecto de cómo proceder en relación con las normas reglamentarias emanadas del Ministerio de Educación y otros organismos.

3. La tecnología de las pruebas basadas en dominios constituye una alternativa la cual vincula el proceso de evaluación del rendimiento estudiantil al proceso de instrucción, lográndose con esto una evaluación más válida que la que se practica cuando ambos procesos marchan separadamente, sin que exista entre ellos una relación real. En efecto, la utilización adecuada y sistemática de pruebas basadas en dominios es consustancial con un proceso instruccional diseñado y desarrollado según criterios técnicos fundamentados en la investigación. Esto a su vez responde a las exigencias de una docencia ejercida por personas preparadas científica y técnicamente para tales funciones.

4. Aunque los avances experimentados por la tecnología de las pruebas basadas en dominios la han colocado en un nivel de aparente complejidad y difícil comprensión para el usuario no experto, la utilización de esta tecnología en Venezuela requiere, como se dijo en otro trabajo (Salcedo, 1980), un tratamiento "al alcance del docente común, sacrificando tecnicismos de difícil comprensión a procedimientos de fácil aplicación y evidente utilidad" (p. 28). En este sentido, la factibilidad de la utilización adecuada de estas pruebas se incrementa si se la asocia a la creación de departamentos y unidades de evaluación en los planteles e instituciones que los requieran, entre cuyas funciones básicas estarían las relacionadas con la elaboración de pruebas y otros instrumentos, su validación y la conformación de bancos de ítems para diferentes asignaturas y niveles escolares.

5. La variedad de métodos y procedimientos considerados en este trabajo respecto de validez, confiabilidad y establecimiento de niveles de rendimiento o estándares conduce a reflexionar acerca de aquéllos que se adecúen mejor a la realidad venezolana. En este sentido, los estudios de postgrado en Educación, y particularmente en Evaluación, constituyen un contexto adecuado para propiciar y estimular la investigación en las áreas analizadas. Esto permitiría la formulación, reformulación y/o adaptación de métodos y técnicas a las condiciones reales de la educación del país.

REFERENCIAS

Angoff, W.H. (1971). Scales, norms, and equivalent scores. En R.L. Thorndike (Ed.), **Educational measurement** (pp. 508–600). Washington, D.C.: American Council on Education.

Baker, E.L. & Herman, J.L. (1983). Task structure design: Beyond linkage. **Journal of Educational Measurement, 20** (2), 149–163.

Block, J.H. (1978). Standards and criteria: A response. **Journal of Educational Measurement, 15** (4), 291–295.

Burry, J., Herman, J.L., & Baker, E.L. (1984). **A practical approach to local test development.** Resource paper No. 6. Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.

Burton, N.W. (1978). Societal standards. **Journal of Educational Messurement, 15** (4), 263–271.

Carver, R.P. (1970). Special problems in measuring change with psychometric devices. En **Evaluative research: Strategies and methods.** (pp. 48–63). Pittsburgh, Pa.: American Institutes for Research.

Cronbach, J.L. (1971). Test validation. En R.L. Thorndike (Ed.). **Educational Measurement,** (pp. 443–507). Washington, D.C.: American Council on Education.

Cronbach, J.L., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). **The dependability of behavioral measurements: Theory of generalizability for scores and profiles.** New York: John Wiley.

Ebel, R.L. (1972). **Essentials of educational measurement.** Englewood Cliffs, N.J.: Prentice–Hall.

Glass, G.V. (1978). Standards and criteria. **Journal of Educational Measurement, 15** (4), 237–261.

Haladyna, T.H. & Roid, G.H. (1983). Reviewing criterion referenced test items. **Educational Technology, 23** (8), 35–38.

Hambleton, R.K. (1978). On the use of cut–off scores with criterion referenced tests in instructional settings. **Journal of Educational Measurement, 15** (4), 277–290.

Hambleton, R.K. (1982). Test score validity and standard–setting methods. En R.A. Berk (Ed.), **Criterion–referenced measurement**

The state of the art. (pp. 80–123). Baltimore, MD: The johns Hopkins University Press.

Hambleton, R.K. & Eignor, D.R. (1980). Competency test development, validation, and standard setting. En R.M. Jaeger & C.K. Tittle (Eds.), **Minimum competency achievement testing.** (pp. 367– 396). Berkeley, CA: McCutchan Publishing Corporation.

Hambleton, R.K. & Novik, M.R. (1973). Toward an integration of theory and method for criterion–referenced tests. **Journal of Educational Measurement, 10,** 159–170.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion–referenced testing and measurement: a review of technical issues and developments. **Review of Educational Research, 48** (1), 1–47.

Harris, C.A. (1972). An interpretation of Livingston's reliability coefficient for criterion–referenced tests. **Journal of Educational Measurement, 9** (1), 27–29.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. (1973) **Domain–referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project.** Monograph Series in Evaluation No. 1, Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.

Huynh, H. (1976). On the reliability of decisions in domain–referenced testing. **Journal of Educational Measurement, 13** (4), 253–264.

Linn, R.L. (1978). Demands, cautions, and suggestions for setting standards. **Journal of Educational Measurement, 15** (4), 301–308.

Linn, R.L. (1979). Issues of validity in measurement for competency based programs. En M.A. Bunda & J.R. Sanders (Eds.), **Practices**

and problems in competency–based measurement (108–123). Washington, D.C.: National Council on Measurement in Education.

Linn, R.L. (1980). Issues of validity for criterion–referenced measures. **Applied Psychological Measurement, 4** (4), 547–561.

Marshall, J.L. & Haertel, E.H. (1976). **The mean split–half coefficient of agreement: A single administration index of realibility for mastery tests.** Manuscript, University of Wisconsin.

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

Millman, J. (1974). Criterion–referenced measurement, En W.J. Popham (Ed.), **Evaluation in education: Current applications** (pp. 309 397). Berkeley, CA: McCutchan Publishing Corporation.

Nedelski, L. (1954). Absolute grading standards for objective tests. **Educational and Psychological Measurement, 14,** 3–19.

Popham, W.J. (1975). **Educational evaluation.** Englewood Cliffs, N.J. Prentice–Hall.

Popham, W.J. (1978a). As always, provocative. **Journal of Educational Measurement, 15** (4), 297–300.

Popham, W.J. (1978b). **Criterion–referenced measurement.** Englewood Cliffs, N.J.: Prentice–Hall.

Popham, W.J. (1981). Modern educational measurement. Englewood Cliffs, N.J.: Prentice–Hall.

Popham, W.J. & Husek, T.R. (1969). Implications of criterion–referenced measurement. **Journal of Educational Measurement, 6** (1), 1–9.