Psicología, tercera época ▪ Revista digital arbitrada
Vol. 32, Núm. 1-2013 | pp. 75-105 | ISSN: 1316- 0923

Construcción de pruebas criteriales: aspectos teóricos y metodológicos

Luisa E. Lezama
Escuela de Psicología, Universidad Central de Venezuela
luisalezama@yahoo.com

Resumen

Durante mucho tiempo la conceptualización y diseño de las pruebas criteriales o tests referidos a un criterio, estuvieron regidos por los mismos procedimientos y análisis estadísticos que las pruebas normativas; no obstante, se debe reconocer en atención a la naturaleza del numeral que ellas generan, que cada una requiere caminos particulares para garantizar las propiedades psicométricas de confiabilidad y validez. Partiendo de ese reconocimiento, este artículo expone detalladamente la construcción de las pruebas criteriales, precisando las actividades, los diseños de investigación y los procesos estadísticos más usados en cada una de las cinco etapas comprometidas en la elaboración de tests estandarizados. Finalmente, se permite identificar diferencias y similitudes puntuales entre pruebas normativas y criteriales.

Palabras clave: pruebas criteriales, pruebas normativas, etapas en la construcción de test estandarizados, confiabilidad, validez.

Recibido: 16 de julio de 2012
Aceptado: 15 de octubre de 2012
Publicado: 05 de junio de 2013

Psicología ▪ Refereed journal
Volume 32, Issue 1-2013 | Pages 75-105 | ISSN: 1316- 0923

Construction of criterion-referenced tests: Theoretical and methodological aspects

Luisa E. Lezama
Escuela de Psicología, Universidad Central de Venezuela
luisalezama@yahoo.com

Abstract

For a long time the conceptualization and design of criterion-referenced tests were ruled by the same procedures and statistical analysis as norm-referenced tests. However, taking into account the nature of the numeral they generate, it must be recognized that each one requires particular ways to ensure the psychometric properties of reliability and validity. Based on this recognition, this article describes in detail the construction of criterion-referenced tests, determining the most used specific activities, research designs and statistical processes in each of the five stages involved in building standardized tests. Finally, we can identify specific differences and similarities between norm and criterion-referenced tests.

Key words: criterion-referenced tests, norm-referenced tests, steps in the construction of standardized tests, reliability, validity.

Received: Jul. 16, 2012
Accepted: Oct. 15, 2012
Published: Jun. 05, 2013

Conceptualización

Un instrumento estandarizado de medición busca asignar un numeral a determinadas propiedades de las personas con atención a ciertas reglas. Dicho numeral provee una información cuya naturaleza determina el tipo de prueba que se trate. De este modo, si el número asignado a una persona adquiere significado cuando se le analiza en función del área examinada y permite hacer descripciones de esa persona en atención a tal área, se está en presencia de una prueba *referida a un criterio*; pero, si el número asignado a un sujeto toma sentido cuando se le compara con el desempeño de las otras personas de su grupo, debemos reconocernos en presencia de una *prueba normativa*.

Como apreciamos, el modo de interpretar el número asignado por los instrumentos estandarizados permite clasificarlos como *pruebas criteriales o como pruebas normativas*, cada una de las cuales sigue procesos de construcción particulares, siempre con miras a garantizar óptimas propiedades psicométricas. Este artículo trata detalladamente del proceso de diseño y construcción de las pruebas *criteriales o referidas a un criterio*, precisando cada una de las etapas, actividades y procesos estadísticos más usados; finalmente, se permite identificar diferencias puntuales entre *pruebas normativas y criteriales*.

Cuando los asesores, los psicólogos clínicos o los educadores requieren describir intensa y detalladamente las conductas presentes en el repertorio de una persona, o identificar sus déficit de conocimientos, o precisar sus competencias en un área, deberán recurrir a los ya aludidos *tests criteriales o tests referidos a un criterio*, los cuales permiten comparar las capacidades o conductas de una persona con un patrón absoluto que detalla los repertorios conductuales, conocimientos o destrezas que deberían estar presentes y que son inherentes a ese dominio examinado.

Tomemos como ejemplo una prueba criterial hipotética que llamaremos Prueba de matemática de 5to grado (PM5) y supongamos que ella mide, entre otros contenidos, operaciones matemáticas básicas y fracciones. Como se trata de una prueba referida a un criterio, el número que se le asigne a un examinado permitirá describir sus competencias en esa área. De ahí que si un chico califica con 10 puntos el examinador, podría decir que él es capaz de resolver suma, resta y multiplicación de fracciones, pero no maneja lo relativo a multiplicación de fracciones, que es lo que previamente se había

determinado para ese puntaje. Con lo anterior se quiere significar que con la prueba referida a un criterio permite describir al evaluado en función de un área o dominio particular, identificando lo que está en su repertorio y lo que no.

Las *pruebas criteriales* surgieron como una alternativa al uso inapropiado de las *pruebas normativas* dentro de los contextos académicos, ya que estas últimas indicaban, por ejemplo, que el estudiante *X* superaba al estudiante *Y* o que el estudiante *Z* fue el mejor de todos los que tomaron la evaluación; no obstante, dentro de ese contexto esa información es absolutamente irrelevante. En su momento, los educadores reconocieron que su interés era poder describir qué contenidos dominaba o no el estudiante *Y*; es decir, determinar qué sabía, por ejemplo, resolver operaciones básicas con números enteros pero no con fracciones, posibilidad que ofrecen las *pruebas criteriales* y no *las normativas,* que eran las que originalmente se empleaban.

Debido a ese punto de partida, durante algún tiempo *las pruebas referidas a un criterio* fueron diseñadas y evaluadas usando los patrones característicos de los *tests normativos*; no obstante, los desarrollos que tomaron lugar durante la década de los sesenta y setenta delimitaron procedimientos que permiten conducir análisis de ítems, preparar cuadros de conversión y efectuar estudios de confiabilidad y validez característicos de las *pruebas criteriales*, independientes de los estudios llevados a cabo para las *pruebas normativas*. Dichos procedimientos son los mismos que se emplean en la actualidad, por lo que frecuentemente nos encontraremos reportando literatura de esas décadas, en tanto vigentes hoy día.

Con la evolución del campo de las pruebas a lo largo de décadas, los *instrumentos criteriales*, que se iniciaron como una extensión de los *normativos* dentro de los salones de clases, han alcanzado un desarrollo importante, tanto que en países como Alemania, Bélgica, Japón, etc., en el marco del conocido Programa Internacional de Evaluación de Estudiantes (PISA), se usan para examinar el rendimiento de los alumnos con miras a efectuar su valoración internacional; también son empleadas en el sistema educativo básico y medio mexicano, no solo para evaluar el aprovechamiento de los estudiantes, sino para recabar información acerca de la eficiencia del propio Sistema Educativo Nacional (Backoff, Sánchez, Peón, Monroy y Tanamachi, 2006).

Adicionalmente debe destacarse que las *pruebas referidas a un criterio* salieron de los salones de clases y actualmente se emplean en diversidad de contextos, siempre que se requiera caracterizar a una persona en atención a un dominio particular; así, por ejemplo, se puede citar el Intermittent Explosive Disorder-Revised (IED-R), desarrollado por Coccaro, Kavoussi, Berman y Lish (1998), que representa un instrumento que contribuye al diagnóstico del trastorno de explosivo intermitente; y también la Prueba de Ideación Suicida de Bravo de Cardozo y Garbán (1991), instrumento criterial que informa la frecuencia e intensidad de los pensamientos suicidas.

Independientemente de si la *prueba referida a un criterio* examinan conocimiento de matemáticas, trastorno de ánimo o ideación suicida, es usual notar que los ítems que la componen son muy similares a la conducta que intentan medir; así, por ejemplo, si se trata de un instrumento que examina el trastorno explosivo intermitente, podrá apreciarse que los ítems tendrán que ver con manifestaciones repentinas de ira intensa, reacciones desproporcionadas ante estímulos leves, entre otras. En los casos en que la variable examinada resulte más abstractas que las anteriormente ejemplificadas, como cuando se evalúa la capacidad de análisis y síntesis de hechos históricos, se requiere hacer esfuerzos especiales para definir el dominio y garantizar que los ítems sean una *muestra* representativa del mismo.

El anterior es un punto clave dentro de la caracterización de los *tests criteriales,* por lo que sus constructores son especialmente exigentes con la delimitación del área a examinar (universo) y luego con la elaboración de ítems que lo representen (muestra). En tanto los reactivos sean representativos del universo medido, el usuario podrá preparar una descripción detallada de las conductas que se encuentran en el repertorio del examinado y precisar las ausentes; igualmente podrá identificar lo que puede o no hacer; o lo que sabe y lo que no.

Construcción de una prueba criterial

Tavella (1978) señala que hay cinco etapas en la construcción de un instrumento de medición; a saber: *preparatoria, exploratoria, experimental, definitiva y de revisión*. Si bien estas etapas son aludidas frecuentemente al hablar de tests referidos a normas, resultan útiles durante la preparación de cualquier otro tipo de test, ya que proveen al constructor de un esquema

de trabajo que deja claro qué actividades realizar en cada una de ellas y cuáles son los estándares de calidad que deben ser alcanzados.

El autor acota que si dentro de una etapa las actividades son adelantadas con éxito, se puede pasar a la siguiente, pero en caso de que los indicadores no den cuenta de calidad se requerirá regresar a la etapa anterior o tantas etapas hacia atrás, como sea necesario, para alcanzar niveles adecuados de calidad. Seguidamente se detallan las actividades a desarrollar en cada una de ellas.

I. Etapa preparatoria

En esta se deben alcanzar varios objetivos, entre los cuales destacan: declarar la población a la que se dirigirá la evaluación; precisar la finalidad del test, es decir, si se empleará para hacer una descripción o una estimación; preparar el esquema descriptivo; decidir el formato de la prueba; redactar los ítems y elaborar las instrucciones de administración y calificación.

Cuando la prueba tiene fines predictivos, es necesario agregar una actividad más a las anteriormente referidas, que consiste en analizar sistemáticamente la variable a estimar. De este modo, si se quiere predecir el éxito de un vendedor se podría considerar el número de pólizas vendidas mensualmente, o el número de personas que reportan que se sintieron bien atendidas por él, etc. hasta seleccionar el mejor indicador de la conducta a estimar; posteriormente, se procede a identificar una variable que pueda anticipar ese éxito (predictor) y a construir el instrumento que permita medirlo.

Ya sea que la *prueba criterial* se emplee para hacer descripciones detalladas de un dominio o para estimar, el constructor debe preparar un *esquema descriptivo o tabla de especificaciones,* que es un esqueleto que sistematiza el área de comportamiento a evaluar. Para prepararlo recurre a una variedad de fuentes de información que permiten delimitar el universo a medir; entre ellas usa libros de textos, resúmenes, consulta con expertos, entrevista con las personas en la que efectuará la medición, observación sistemática, etc.

Así, para diseñar el *esquema descriptivo* de la PM5, el constructor podría recurrir a diversas fuentes de información (programa oficiales del Ministerio, profesores, libros de 5to grado, etc.) e identificar los contenidos de esa materia (operaciones matemáticas básicas con números enteros y fracciones) y los objetivos que se deben alcanzar en ella (conozca los términos, resuelva operaciones, aplique las operaciones en situaciones de la vida cotidiana), y generar un esquema como el mostrado en el cuadro 1.

Cuadro 1
Esquema descriptivo de la PM5

Objetivos	**Contenidos**	
	Números enteros	Fracciones
Conoce términos	2	3
Resuelve sumas	3	2
Resuelve restas	3	2
Resuelve multiplicaciones	3	2
Resuelve divisiones	3	2
Aplica a la vida cotidiana	3	2
Total ítems	17	13

El *esquema descriptivo* cumple una doble función; por una parte, especifica a los redactores de ítems qué reactivos deben elaborar, de modo de cubrir apropiadamente toda el área examinada, evitando sobremedir un tópico y dejar otros sin examinar; y por otro lado, comunica a los usuarios del instrumento qué es lo que este mide. Así, el *esquema descriptivo* de la PM5 evalúa el conocimiento acerca de las operaciones básicas con números enteros y fraccionarios, pero no examina los conocimientos de números negativos, ni de potencia; el usuario conoce con precisión qué se examina y qué no con ese instrumento y consecuentemente qué puede decir y que no acerca de los examinados.

Una vez bosquejado un primer *esquema descriptivo,* se solicita a expertos en el tópico examinado que evalúen la correspondencia entre este y el área a medir. Si es necesario, el constructor realiza los ajustes señalados por los expertos y continúa su proceso sobre bases más firmes, procediendo entonces a seleccionar el formato de la prueba y decidir el número y tipo de ítems pertinentes.

De este modo, debe seleccionar un *formato de prueba* cónsono con el dominio medido (personalidad o rendimiento, etc.) y con las características de las personas a quienes va dirigido (niños, adultos, personas con discapacidad visual, etc.), y al mismo tiempo debe decidir si la prueba será de papel y lápiz, administrada por computadora, de ejecución u oral, si será de administración individual o colectiva o si deberá corregirse de forma objetiva o semiobjetiva, manual o por computadora, de modo de verificar que cuenta con los recursos humanos y materiales para tales fines.

La decisión en torno al *número de ítems* de la prueba no es sencilla. El constructor debe tener en cuenta tres aspectos esenciales: 1. El nivel mínimo de habilidad exigida al examinado, de modo que si la prueba demanda un rendimiento de 95% de respuestas correctas, debe tener menos reactivos que si exige solo un desempeño de 50% de ítems correctos (Dembo, 1983); 2. El nivel estimado de funcionamiento de los examinados; cuando se suponga que tal nivel de funcionamiento es alto, se deben emplear más preguntas que cuando se estime que es bajo; y 3. La proporción de pérdida, que alude a dos tipos de errores: a) decir que una persona posee una destreza o conducta cuando no la posee; o b) decir que no la tiene cuando efectivamente está en su repertorio. Mientras mayor sea la probabilidad de cometer estos tipos de errores, mayor deberá ser el número de reactivos incluidos en el test. Todo esto se conversa con expertos durante la construcción del esquema descriptivo de prueba, para recibir su opinión acerca de este tópico.

En cualquier caso, el error más importante que hay que evitar cometer con relación a la longitud de una prueba criterial es el de "demasiado pocos". Popham (1978) recomienda, con base en su experiencia, medir cada tópico con un aproximado entre 10 y 20 reactivos, con lo cual se evita el error obvio de usar unos pocos planteamientos, para conocer el estatus del examinado respecto a un área de comportamiento.

En relación con el *tipo de reactivo* a elaborar, el constructor debe considerar los procesos mentales que requiera elicitar; tiene para escoger entre ítems de respuesta breve, verdadero falso, elección múltiple, pictóricos, de apareamiento o de ensayo. Puede, por ejemplo, recurrir a un ítem de tipo respuesta breve si debe medir el recuerdo o la memoria; a reactivos pictóricos, si se quiere conocer la capacidad para transformar información de prosa a gráfico, o viceversa; a planteamientos de verdadero falso, si quiere saber si una gama de conductas está dentro del repertorio de un examinado; o a elementos de elección múltiple, si la persona debe efectuar algún proceso de evaluación y síntesis.

La redacción de ítems de un test criterial se hace siguiendo las especificaciones del esquema descriptivo; es decir, se preparan los reactivos señalados en dicho esquema, lo cual permite garantizar la producción de un conjunto de elementos que representan legítimamente el área del comportamiento medida, caso en el cual se dice que los reactivos poseen *homogeneidad derivativa*. Como se puede apreciar, no hablamos de una *homogeneidad funcional* en el sentido de que los examinados deben contestar todos los planteamientos

correcta o incorrectamente, lo que se está diciendo es que los reactivos serán homogéneos en tanto sean congruentes con las especificaciones del test.

Otra actividad que se lleva a cabo durante la *etapa preparatoria* es la redacción de las instrucciones para los examinados y los calificadores. En las instrucciones para los examinados se les indica, en un lenguaje claro y sencillo, el tipo de ejecución que se requiere de ellos, el tiempo que podrían tardar en contestar y el modo de registrar sus respuestas; en las instrucciones para los calificadores, se detalla la forma de corregir, si habrá o no bonificaciones, cuándo descontinuar la administración, etc. Las instrucciones redactadas deberán ser consideradas a la luz de su funcionamiento en la práctica, con lo cual se considerarán en las etapas subsiguientes de la construcción de la prueba.

Una vez que se han llevado a cabo estas actividades, el constructor dispone de un instrumento completo que posee homogeneidad derivativa, un formato de prueba congruente con la variable examinada y la población objeto de evaluación, conformada por ítems ajustados al proceso de interés, lo que hace posible avanzar al momento siguiente.

II. Etapa exploratoria

El objetivo fundamental que persigue el constructor de una prueba criterial durante la *etapa exploratoria*, es evaluar la calidad de los reactivos preparados en la etapa anterior, para lo cual los somete a dos tipos de análisis, uno cualitativo, regularmente llamado *a priori,* y otro cuantitativo, también denominado *a posteriori*. Adicionalmente, durante esta etapa se precisa la duración de la prueba y se evalúa el funcionamiento de las instrucciones para la administración y la calificación del instrumento.

1. Procedimiento a priori

Para efectuar este análisis se recurre, una vez más, a expertos en el área examinada, quienes deben evaluar la *congruencia* de los ítems con el *esquema descriptivo*. Cada experto deberá señalar si una premisa es congruente o no con sus especificaciones y en los casos de incongruencias deberá argumentar tal categorización. Aquellos reactivos considerados congruentes por todos los expertos, van directamente a la prueba, pero cuando dos o más expertos lo identifican como incongruente y señalan las mismas razones, hay evidencias suficientes para descartarlo; si al menos un experto hace algún señalamiento

en torno a un reactivo, este deberá ser mejorado en el sentido señalado por ese experto.

Siempre es más trabajoso categorizar un ítem como incongruente, ya que hay que argumentar los motivos para tal juicio. Algunos expertos pueden suavizar su opinión para ahorrarse trabajo; por tal motivo, el constructor de prueba debe desarrollar estrategias que le permitan precisar la calidad del juicio de los expertos como, por ejemplo, intercalar 3 planteamientos incongruentes por cada 25 a evaluar, esperando que los expertos detecten la incongruencia con las especificaciones, en una suerte de escala de sinceridad. Las opiniones de aquellos expertos que omitan dos o más ítems expresamente incongruentes, serán totalmente descartadas.

2. Procedimiento a posteriori

Una vez que los reactivos han sido sometidos a un análisis de congruencia, el constructor de la prueba efectúa los ajustes correspondientes y procede al llamado *análisis a posteriori*, el cual consiste en una evaluación estadística de los ítems y de la prueba total. Es así como la prueba se administra a un *grupo de criterio,* seleccionado de un modo absolutamente intencional, nunca azaroso; es decir, si la prueba realiza diagnósticos de lesiones cerebrales, la *muestra de criterio* debe estar conformada por personas que hayan sufrido un accidente cerebrovascular o algún otro tipo de afectación; si la prueba examina matemática de 5to grado, se buscarán niños que hayan cursado matemática de 5to grado. Para efectuar tales análisis estadísticos, el constructor recurre a uno de dos tipos de *diseños de investigación:*

a) Diseño de dos grupos o de grupos contrastados

Se seleccionan intencionalmente los miembros de dos grupos, de modo que los integrantes de uno de ellos posean la habilidad o característica examinada por la prueba y los participantes del otro no la posean. Bajo igualdad de condiciones se administra la prueba criterial a los dos grupos contrastados y se somete los datos a un análisis estadístico, esperando que la mayoría de las personas que posee la habilidad haga bien cada reactivo o muestre la característica y quienes no poseen tal habilidad lo hagan mal o no la evidencien. Para efectuar un análisis a posteriori de la PM5 que aludimos antes, su constructor podría seleccionar un grupo que haya cursado quinto grado y otro que esté en tercero, y a ambos administrarles la PM5. Si el instrumento efectivamente examina el aprovechamiento de matemáticas en el grado, los alumnos de quinto deberían mostrar una excelente ejecución, la

cual contrastaría significativamente con el comportamiento de los alumnos tercer grado.

b) Diseño de pre-postest

En este caso se selecciona un solo grupo de criterio, al cual le administra la prueba criterial; luego se somete a alguna manipulación efectiva de la variable medida por el instrumento y posteriormente aplica la misma o su forma paralela. Se espera que si la prueba criterial mide el dominio especificado, las personas obtengan una deficiente ejecución en el pretest que contraste con su buena ejecución en el postest. En el ejemplo de la prueba de matemática para 5to grado, que se viene comentando, el constructor podría proceder administrando la prueba el primer día del año escolar y luego cuando este haya finalizado; se espera que en la primera aplicación el instrumento revele una pobre ejecución por parte de los examinados y un alto rendimiento en el postest.

Independientemente del diseño empleado (dos grupos o pre-postest), el constructor de pruebas deberá considerar las variables extrañas que podrían afectar su investigación, tales como la maduración, la adivinación, la memoria, el mismo test, etc. e introducir los mecanismos de control de variables extrañas que pudieran afectar los datos.

Análisis estadísticos

Una vez obtenidos los datos, estos se deben procesar estadísticamente para dar cuenta de la *dificultad* y de la *capacidad discriminativa* de los ítems que conforma el instrumento.

1. Dificultad del ítem

La dificultad del ítem (p) se obtiene al contar el número de personas que lo contestan correctamente *(f)* y dividirlo entre el número de personas del grupo *(N)* $p = f/N$. Se expresa en una escala de proporción, por lo cual asume valores entre 0 y 1, de modo que mientras más cercano a 1, más fácil será el reactivo, mientras más cercano a 0, más difícil resultará; así, un planteamiento cuya $p = 0{,}60$ informa que 60% de las personas del grupo lo contestó correctamente.

Si se retoma el ejemplo de la PM5 y se plantea a la hora de hacer el análisis a posteriori, aplicar un diseño de pre-postest, la evaluación de los reactivos que

consideran su nivel de dificultad llevará a categorizar como un "buen ítem" aquel que arroje: un nivel de dificultad cercano a 0 en el pretest, cercano a 1 en el postest y revele diferencias significativas entre el pre y el postest. Un reactivo será categorizado como "inapropiado", por ejemplo, si resulta fácil en el pretest (*p* cercana a 1*),* o difícil en postest *(p* cercana a 0*)* o con niveles de dificultad similares en pre y postest.

En síntesis, se puede afirmar que, en términos del nivel de dificultad, para una prueba criterial será "bueno" aquel reactivo que resulte muy difícil en el pretest o para el grupo sin la habilidad y al mismo tiempo fácil en el postest o para el grupo con la habilidad, y que demuestre diferencias estadísticamente significativas entre las condiciones comparadas.

2. Capacidad discriminativa del ítem

Según Berk (1980), la c*apacidad discriminativa de un ítem* de una prueba referida a un criterio se puede estimar a partir de los *métodos de Cox y Vargas* y de *Klein y Kosekoff*, que están sujetos al diseño de investigación empleado.

a) Método de Cox y Vargas

También denominado método de sensibilidad instruccional, se aplica en los diseños de pre y postest y asume valores entre +1 y -1. El constructor de pruebas calcula el índice de discriminación de cada pregunta, restando a la proporción de personas que lo contesta correctamente en el postest, la proporción de personas que lo contesta correctamente en el pretest.

El cuadro 2 muestra que el ítem 1 revela la mayor capacidad discriminativa entre los tres reactivos del ejemplo, ya que es contestado correctamente por todas las personas en el postest e incorrectamente por todos en el pretest. Esto es precisamente lo que se busca de un reactivo en una prueba criterial: que sea capaz de diferenciar entre aquellos que poseen el dominio o el rasgo considerado, de aquellos que no los poseen.

Cuadro 2
Índice de discriminación de Cox y Vargas

N° del ítem	p en postest	p en pretest	Cox y Vargas
1	1	0	+1
2	0	1	-1
3	0,20	0,20	0

El reactivo 2, en cambio, es contestado bien por todas las personas en el pretest, pero mal por todos en el postest, lo cual es completamente opuesto a lo que se busca de un ítem en un test criterial, donde después de recibir el adiestramiento es que se debe observar buena ejecución en los reactivos y no antes; ello hace de la 2 una pregunta descartable. Los planteamientos que muestran el mismo comportamiento, tanto en el pre como en el postest, también deben ser eliminados, puesto que no aportan capacidad discriminativa a la prueba, lo cual se puede observar en el ítem 3, que fue contestado correctamente tanto en el pre como en el postest por 20% de las personas, resultando cero su capacidad discriminativa.

En síntesis, mientras más se acerque a +1 el índice de *Cox y Vargas,* mayor será la capacidad discriminativa de los ítems de una prueba criterial, cuyo *análisis a posteriori* se basa en un diseño de pre-posprueba.

b) Método de Klein y Kosekoff

Se aplica en los diseños de *dos grupos o grupos contrastados* y asume valores entre +1 y -1. El constructor de pruebas calcula el índice de discriminación de cada ítem, restando a la proporción de personas que lo contesta correctamente en el grupo que sabe posee el dominio o la característica considerada, la proporción de personas que lo contesta correctamente en el grupo que sabe *no posee* el dominio. El tipo de análisis es análogo al recientemente detallado para el *método de Cox y Vargas*.

Un constructor de pruebas criteriales, mientras lleva a cabo el análisis estadístico de los reactivos, debe mantener presente el esquema descriptivo que elaboró para, en función de él y recurriendo a los análisis tanto *a priori* como *a posteriori,* decidir cuáles planteamientos son definitivamente aceptados o descartados del instrumento. Será una medida inapropiada descartar "ítems cualitativamente congruentes" porque no cumplieron todos los estándares estadísticos; también será una medida equivocada descartar todos los reactivos que miden un mismo aspecto sin reemplazarlos por otros de mejor calidad, ya que se estaría dejando de cumplir con las especificaciones de la prueba.

Cuando se han construido reactivos de elección múltiple, es apropiado analizar el comportamiento de las opciones, del mismo modo como se acaba de reseñar que se examinan las respuestas correctas de los ítems. En los casos que tal análisis resulte insuficiente, es recomendable adelantar un *análisis de procesos,* en el cual se solicita a algunas personas del grupo de criterio que

relaten los pasos que le llevaron a seleccionar una determinar opción. Este bagaje de información permite al constructor de pruebas tomar decisiones en torno a la calidad de un reactivo.

Después de realizar el análisis de los reactivos siguiendo los procedimientos *a priori y a posteriori,* el constructor debe tomar decisiones en torno al funcionamiento de las instrucciones de administración y corrección de la prueba, e igualmente precisar el tiempo de administración que demanda el instrumento, según lo que haya ocurrido cuando la administró a los participantes.

III. Etapa experimental

Cuando los reactivos que componen el instrumento han sido analizados y ajustados según las directrices de los análisis cuali y cuantitativos, cuando ya se han optimizado las instrucciones, y cuando se tiene precisión en torno a la duración de la prueba, esta se administra nuevamente a una muestra intencional con la misión de verificar el óptimo funcionamiento de los reactivos y del instrumento como totalidad. Si se observa que los ajustes han sido adecuados, se pasa a la siguiente etapa.

IV. Etapa definitiva

Esta etapa toma lugar cuando el instrumento en construcción tiene tal nivel en sus ítems, que puede procederse a efectuar los estudios requeridos para dotarlo de las propiedades de *confiabilidad y validez*, así como para elaborar los *cuadros de conversión pertinentes*. Si bien es cierto que en términos formales es en esta etapa que se realizan estos estudios y son los datos derivados de ellos los que se registran en el manual de la prueba, no es menos cierto que los resultados a obtener acá están determinados por los resultados tanto de la *etapa exploratoria* como la *experimental*.

1. Confiabilidad

La confiabilidad de una prueba criterial alude a la *consistencia o repetibilidad* con la cual se puede describir el comportamiento de una persona en relación con ese dominio conductual. El constructor de este tipo de pruebas puede dar cuenta de tal consistencia o repetibilidad en distintos sentidos: determinar *cuán estables en el tiempo* son las decisiones que la prueba ayuda a tomar; precisar *cuán equivalentes* son dos pruebas paralelas a la hora de

tomar decisiones; verificar si la prueba es *internamente consistente;* conocer el a*cuerdo entre observadores* necesario cuando el instrumento sea de calificación semiobjetiva. Seguidamente cada uno de ellos.

a) Test retest

Cuando se necesitan evidencias de que las descripciones y las decisiones que la prueba ayuda a tomar un día, son las mismas que las descripciones y las decisiones tomadas dentro de un tiempo, se hace indispensable efectuar estudios de *estabilidad*.

El esquema tradicional de trabajo para las pruebas normativas, en situaciones como estas, consiste en administrar el instrumento a un grupo, dejar pasar un tiempo y administrarlo de nuevo bajo idénticas condiciones, para luego proceder a determinar una correlación, generalmente *producto momento de Pearson;* no obstante, dicho procedimiento demanda condiciones especiales de varianza que no se presentan cuando la prueba diseñada es del tipo criterial. Para superar tal limitación dentro del contexto de las pruebas referidas a un criterio, se ha propuesto el cálculo de una variedad de índices, entre los cuales Almerich y Bo Bonet (2006) refieren las propuestas de Hambleton y Novick, Livingston, Berk y Subkoviak, y muy especialmente se destaca el procedimiento usado por Popham.

Popham (1978) propone un esquema de trabajo parsimonioso a la hora de determinar la estabilidad temporal de las decisiones tomadas. En tanto la idea es responder la pregunta: ¿Es consistente este instrumento al clasificar a las personas en una de dos categorías (aceptados y rechazados) o más categorías (normal, anorexia subclínica, anorexia)? La respuesta podría conllevar: 1. administrar la prueba (test) y clasificar a las personas en la(s) categoría(s) correspondiente(s); 2. tiempo después repetir el proceso (retest) con las mismas personas; y 3. correr alguna prueba de significación estadística (*Chi cuadrado* χ^2, p. e.). En el caso de que no se rechace la hipótesis nula, se puede afirmar que el instrumento lleva a tomar las mismas decisiones en los dos momentos, lo cual implica que es consistente a la hora de permitir la toma de decisiones.

b) Formas paralelas

Hay ocasiones en las que se requiere más de una forma de un mismo test como, por ejemplo, cuando se aplica una misma prueba con mucha frecuencia o cuando se necesita administrar el instrumento en un pre

y postest, pero el intervalo es muy corto para garantizar que la misma prueba no funcione como variable extraña. En tales casos el constructor de pruebas puede recurrir a preparar pruebas paralelas y a "confiabilizarlas" a través de un procedimiento de *equivalencia*.

Para preparar dos pruebas equivalentes se elaboran los reactivos de cada una de las formas a partir del mismo *esquema descriptivo*, para luego adelantar los análisis de congruencia y estadísticos reseñados antes, manteniendo siempre en mente que dichas formas deben poseer *homogeneidad derivativa*. Posteriormente, se administran esas dos pruebas consecutivamente a un grupo de criterio y los datos obtenidos se procesan estadísticamente (*Chi cuadrado* χ^2, p. e.); si no existen diferencias significativas entre los datos derivados de las dos pruebas equivalentes, se implica que sus resultados son repetibles, de lo que se desprende que poseen *confiabilidad de formas paralelas.*

c) Consistencia interna

La consistencia interna de las pruebas referidas a un criterio, queda demostrada cualitativamente durante los estudios *a priori* cuando se determina su homogeneidad derivativa; no obstante, también se puede recurrir al análisis estadístico para informar si un grupo de reactivos es congruente al examinar el dominio que pretende medir; es decir, si por ejemplo, los planteamientos que miden suma de fracciones generan patrones repetibles de respuestas, o si son consistentes entre sí, los que miden multiplicación de número enteros. Para generar un estadístico se puede comparar, usando una prueba χ^2, la frecuencia de las respuestas correctas a cada reactivo con la mediana de su respectiva área. Si todos los ítems miden el mismo dominio, no debería haber diferencias estadísticamente significativas, en cuyo caso se afirmaría que esa área está conformada por reactivos consistentes entre sí.

Hay que ser especialmente cuidadoso cuando se analicen reactivos que midan un dominio heterogéneo, para no sobreexigirles indicadores de homogeneidad, cuando efectivamente el dominio es heterogéneo. Así, cuando los datos revelen heterogeneidad, corresponde al constructor de test analizar cualitativamente los reactivos y contrastarlos con las especificaciones de la prueba, a los fines de detectar si esa heterogeneidad es característica de la variable, caso en el que no se le hacen ajustes; o si es necesario modificar los reactivos.

d) Acuerdo entre calificadores

Informa el grado en que dos observadores concuerdan en la calificación asignada con un instrumento; consecuentemente, su utilidad existe en tanto el instrumento que se está diseñando requiera de una calificación semiobjetiva; es decir, la confiabilidad de acuerdo entre jueces carece de sentido si el instrumento es de calificación objetiva o se corrige de modo automatizado, ya que en ellos nunca habrá desacuerdo en la calificación asignada, siempre que se respeten las condiciones de estandarización.

Para establecer la confiabilidad de un instrumento de calificación semiobjetiva, se solicita a dos o más personas que corrijan el mismo instrumento. Los datos obtenidos pueden ser analizados atendiendo: o a la naturaleza de los datos generados (correlación *producto momento de Pearson, Phi, puntobiserial, o de contingencia, etc*); o a la dimensión relevante de la conducta examinada, como lo pueden ser frecuencia, duración, intensidad (Lacasella, 2000); o las etiquetas asignadas por los evaluadores usando a un procedimiento de *Chi cuadrado,* decisión que tomará el constructor.

Antes de concluir la exposición en relación con la confiabilidad, deben señalarse dos cosas: la primera es que los procedimientos reseñados antes no son excluyentes entre sí; es decir, si para probar la calidad de un instrumento se requiere determinar su confiabilidad de retest y de consistencia interna, ambos procedimientos deber ser llevados a cabo, así como cualquier otra combinación que sea pertinente. La segunda es que el diseñador de pruebas está en el deber de reportar detalladamente en el manual, los procedimientos, muestras de personas y resultados de los estudios de confiabilidad adelantados, con el objetivo de proveer a los eventuales usuarios de recursos para tomar decisiones acerca de la conveniencia de emplear el instrumento en cuestión.

2. Validez

Dado que un test criterial se puede emplear para *describir* un dominio particular o para *estimar* otra variable, su construcción conlleva la implementación de estudios que garanticen que el instrumento sirve a estos fines (Linn, 1980), para lo cual el diseñador puede recurrir a tres tipos de enfoques de validación; a saber, *validez de selección de dominio, validez descriptiva y validez funcional* (Popham, 1978), cada una de las cuales son detalladas de seguida.

a) De selección de dominio

Esta tiene que ver con dar cuenta que el instrumento muestrea adecuadamente el universo a examinar, por lo cual se afirma que alude fundamentalmente a la calidad con la que se delimita el dominio medido, con las especificaciones de la prueba y con la pertinencia de los reactivos preparados.

Aunque este tipo de validez se delimita en esta *etapa definitiva* del diseño de un instrumento, su constructor trabaja en ella desde el momento mismo de su conceptualización al: 1. delimitar el dominio a examinar; 2. hacer que paneles de expertos examinen su tabla de especificaciones; 3. preparar los reactivos en función de ella; 4. solicitar que otro panel de expertos evalúe la congruencia de los ítems en relación con las especificaciones; y 5. dar preponderancia a las evaluaciones cualitativas sobre las cuantitativas de los reactivos.

El constructor de prueba debe recurrir a recursos como la revisión de libros de textos, resúmenes, programas oficiales de las materias, manuales psiquiátricos, hallazgos empíricos, observación directa, sesiones de entrevistas, etc., con el fin de documentarse en el área que examinará. Adicionalmente, debe seleccionar apropiadamente al panel de expertos que considerarán las especificaciones de la prueba e implementarán los juicios de congruencia de los ítems; mientras más numerosos y calificados académica y experiencialmente sean, mayores garantías se tendrá de que la prueba posee v*alidez de selección de dominio*.

Si bien no es lo usual, también se pueden obtener evidencias numéricas de validez de selección de dominio, correlacionando este instrumento criterial con otro que se sepa mide el mismo dominio, lo cual es un procedimiento muy similar a la validez de constructo de las pruebas normativas, pero dado que tiene poco sentido construir un instrumento para describir exactamente el mismo dominio, esta estrategia es poco empleada, amén del asunto metodológico que tienen que ver con el análisis de la varianza característico de las pruebas referidas a un dominio.

b) Validez descriptiva

Informa qué significa; es la puntuación de una persona en la prueba, al proveer de una descripción clara del dominio conductual que mide el test. Si el instrumento no permite hacer una descripción exhaustiva de la conducta, conocimiento o habilidades cubiertos por él, se le catalogará como no válido.

Para disponer de evidencia empírica se llevan a cabo los llamados *juicios de validez descriptiva*, consistentes en estudios en los que se solicita a usuarios potenciales del instrumento que elaboren reactivos según el *esquema descriptivo* del test y posteriormente se pide a otras personas que determinen la congruencia de dichos ítems con las especificaciones del test. Si las especificaciones de la prueba pudieron ser comprendidas claramente por tales usuarios potenciales y ellos fueron capaces de preparar y/o juzgar reactivos congruentes con las especificaciones, se puede concluir que los usuarios de la prueba llegarán a descripciones similares acerca de la ejecución de una persona y que las mismas serán ajustadas al dominio examinado.

Por ejemplo, para informar acerca de la *validez descriptiva* de la PM5, se puede pedir a seis usuarios potenciales del instrumento que preparen cada uno dos reactivos para determinada especificación (multiplicación de fracciones, p. e.); posteriormente esos 12 reactivos se entregan a otras tres personas para que juzguen su homogeneidad. Si la mayoría de los ítems (90% o más) son relacionados con su especificación correspondiente (multiplicación de fracciones), se puede afirmar que cualquier usuario del PM5 será capaz de comprender las especificaciones de la prueba y describir apropiadamente la ejecución de un examinado.

c) Validez funcional

Cuando se dan pasos para preparar la prueba y se llevan a efecto juicios de validez descriptiva, se tiene información de que el instrumento funciona para *describir el dominio*, pero cuando se le emplea para *efectuar estimaciones* hay que exponer evidencias de *validez funcional.*

La estimación que se quiere efectuar con el instrumento determinará el tipo de estudio a realizar. Como quiera, es necesario diseñar el instrumento (predictor), especificar las medidas del criterio, administrar ambos instrumentos y asociar los puntajes obtenidos; si esa asociación es alta y significativa, se puede afirmar que el predictor (prueba criterial) puede usarse para hacer estimaciones en el criterio. En ese sentido, la prueba criterial posee *validez funcional.*

Si, por ejemplo, se quiere emplear la PM5 para predecir el rendimiento en matemáticas de 6to grado, se pueden asociar las calificaciones de un grupo de niños en la PM5, tomadas al concluir el período académico, con sus calificaciones en 6to grado, una vez que han concluido ese año; si esa asociación es significativa, se puede concluir que conociendo el nivel de

aprovechamiento en el dominio medido por el PM5, se puede anticipar el desempeño en el 6to grado.

Un aspecto que se debe mantener presente es que la misión fundamental de un *test referido a criterio,* es hacer descripciones sustanciales de una conducta, habilidad o aprendizaje, por lo que nunca se deberá sacrificar esta misión, en aras de tomar decisiones o hacer estimaciones (Popham, 1978).

3. Preparación de las calificaciones

Al tiempo que el constructor de una *prueba criterial* delimita el dominio a medir, esquematiza una escala para reportar la ejecución en la prueba, la cual le permitirá comunicar el nivel de ejecución de la persona examinada. Para precisar la escala que emplear debe considerar los fines del instrumento: si será empleado para describir la ejecución, se recurrirá al diseño de *escalas relacionadas con el contenido*, pero si se usará para estimar la ejecución en un criterio se recurrirá a *escalas relacionadas con el rendimiento.*

Lo común a ambos grupos de calificaciones es que sus significados se establecen *a priori*, o sea, sin esperar ver cómo se comportan los examinados; es decir, en la medida en que se construyen los planteamientos el diseñador va visualizando lo que significan determinados puntajes, con lo llega a establecer que 10 respuestas afirmativas implican un trastorno de ánimo de tipo distímico, por ejemplo. Aunque los significados de las calificaciones de las pruebas referidas a un criterio se tienen claros desde temprano en el proceso de construcción de la prueba, es menester esperar disponer de los reportes formales de confiabilidad y validez antes de proceder a precisarlos; por tal razón la literatura señala que se deben preparar en *etapa definitiva*.

Escalas relacionadas con el contenido

Estas escalas permiten describir si una persona posee o no y con cuáles características un dominio específico. Entre ellas se encuentran *las calificaciones de corte, el porcentaje de respuestas correctas, las calificaciones estándares de contenido y las escalas de clasificación.*

a) Calificaciones de habilidad o nivel de pase

Cuando el objetivo de la prueba es informar si una persona ha alcanzado o no un nivel determinado de destreza o es posible asignarle una etiqueta, si aprehendió o no las competencias, se emplea como calificación el conocido

el nivel de pase, es decir, el nivel mínimo de calificación aceptado, de modo que si una persona alcanza esa calificación o una más alta que ella, se puede afirmar que posee el dominio del material, de la conducta o de la competencia. El ejemplo típico de nivel de corte es el puntaje 10 en el sistema educativo venezolano, donde quien puntúe 10 o más aprueba, y quien puntúe 9 o menos deberá repetir el curso; o también la calificación de 40 en las pruebas NECAP del programa de educación común de Nueva Inglaterra, donde resultados de 40 y más indican una habilidad por encima del nivel competente, y resultados de menos de 39 indican habilidad por debajo del nivel competente.

Es importante señalar que cuando se trata de *escalas dicotómicas de habilidad*, cualquier calificación por debajo o por encima del nivel de pase tiene el mismo significado. Así, por ejemplo, si en la PM5 el nivel de pase es 30 puntos, una calificación de 10 puntos o de 29 se considerará que posee el mismo significado; y una calificación de 31 puntos tendrá el mismo significado que una de 40. Esta es considerada una limitación de tales puntajes y la razón por la cual suelen emplearse en conjunto con otro tipo de calificaciones.

b) Porcentaje de respuestas correctas

Como se acaba de señalar, cuando se reporta la ejecución en términos de *niveles de pase* o fracaso, se pierde gran cantidad de información, lo que se puede superar informando, adicionalmente, la calificación de la persona según su ubicación en un continuo de destreza o habilidad a través del *porcentaje o proporción de reactivos contestados correctamente,* que se obtiene de: *% de respuestas = número de respuestas correctas/número total de preguntas*, en cada área de contenido.

Cuando se emplea el *porcentaje de respuestas correctas*, el usuario debe tener presente que dos personas pueden obtener la misma calificación (porcentaje de respuestas correctas) en una misma área, aunque hayan contestado correctamente diferentes reactivos, por tal razón se debe examinar con detalle cuáles son esas respuestas correctas y preparar informes particularizados, lo cual es obvio para las *pruebas referidas a un criterio,* que encuentran su esencia en esta particularización.

c) Calificaciones estándar de contenido

Este tipo de calificaciones es empleado en aquellos instrumentos cuyos constructores estandarizan, incluso, la interpretación de las puntuaciones que recogen dichos instrumentos, al indicar el significado de rangos de puntuaciones; es decir, hay una sección en la prueba donde se reporta que obtener puntuaciones entre *X* y *Y* tiene determinado significado, y que obtener puntuaciones entre *Z* y *W* tiene este otro significado, los cuales se detallan extensamente, de modo que el usuario disponga de categorías descriptivas de las características de los examinados. Es de hacer notar que este tipo de calificaciones demanda un riguroso proceso de construcción y selección de reactivos, para poder garantizar que una misma puntuación tenga un único significado, ya que debe provenir de una única combinación de respuesta de los examinados.

Un ejemplo del uso de este tipo calificación se puede apreciar en "El Programa de Evaluación de Extensión 1 de California del Norte", que examina varias áreas de conocimientos, entre ellas la lectura. En este instrumento las calificaciones para alumnos del décimo año oscilan entre 0 y 30 y tienen varios rangos; uno de ellos es el *estándar de contenido* de 26 al 30, que permite afirmar, acerca de un alumno que caiga en ese rango, que "Demuestra destrezas de lectura que van más allá de las exigencias establecidas para el décimo grado en Carolina del Norte. Comprende una variedad de textos informativos, argumentativos y expresivos. Es capaz de evaluar relaciones causa-efecto, problemas-soluciones expresados en textos…". La prueba dispone de una descripción análoga a la reseñada para cada uno de los rangos de *puntajes estándar de contenido.*

d) Escalas de clasificación

Las *escalas de clasificación* representan el modo de presentar las calificaciones en aquellas pruebas que asignan etiquetas a determinados rangos de calificaciones. En cierto sentido, son análogas a las calificaciones *estándar de contenido* al establecer rangos de calificaciones, pero adicionalmente proveen rótulos que le son asignadas a los examinados que caigan en esos rangos. Un ejemplo de este tipo de puntuaciones es las empleadas en instrumentos como la *escala de Conners* (Conners, 1989), utilizada para clasificar a niños con o sin trastorno por déficit de atención con hiperactividad. Mediante dicha escala se pueden asignar tres etiquetas diagnósticas, a saber: trastorno por déficit de atención con hiperactividad, con predominio de déficit de

atención (DA), trastorno por déficit de atención con hiperactividad, con predominio hiperactivo-impulsivo (HI), o trastorno por déficit de atención con hiperactividad, tipo combinado (TC).

Si se analiza con detenimiento es posible percatarse que las *escalas de clasificación* y las *calificaciones estándares de contenido*, en esencia, obedecen a un mismo principio, esto es, una vez obtenida la calificación directa del instrumento, se ubica dentro de un rango que informa el comportamiento característico del examinado. La diferencia esencial quizás estribe en el hecho de que las *calificaciones estándares de contenido* son de uso exclusivo de la medición en contextos académicos, donde la etiqueta diagnóstica es irrelevante y la preponderancia la tienen las fortalezas y debilidades del aprendizaje del estudiante examinado; mientras que en el caso de las *escalas de clasificación* la etiqueta diagnóstica ocupa el lugar central en tanto se busca justamente asignar determinado rótulo al evaluado, a los fines de segmentarlo y separarlo de otros posibles diagnósticos, con lo cual este tipo de calificación suele ser más frecuente en psicología.

Escalas relacionadas con el rendimiento

Hasta ahora han sido consideradas las calificaciones en función del contenido muestreado, ya sea al describir si una persona tiene o no la capacidad o la característica examinada o al precisar la calidad o intensidad con que la posee; no obstante, con este tipo de calificaciones no se pueden hacer *estimaciones* del comportamiento de los evaluados en situaciones diferentes a la prueba o en el futuro.

Cuando este es el interés debe, primero que nada, establecerse la *validez funcional* en los términos recientemente señalados en la sección correspondiente, manteniendo siempre en mente que las variables a predecirse deben expresarse en términos del rendimiento como, por ejemplo, número de piezas armadas, años de permanencia en un empleo, tipos de comportamientos delictivos, entre otros, ya que no debe perderse de vista la naturaleza de las pruebas referidas a un criterio, las cuales buscan esencialmente describir dominios conductuales.

Una vez garantizada la calidad de la prueba criterial para hacer la estimación, se podrá proceder al diseño de las *tablas de expectación* o a los conocidos *niveles predictivos de ejecución*, según la preferencia del constructor.

a) Tabla de expectación

Es el recurso que se emplea en el contexto de las pruebas referidas a un criterio para hacer estimaciones. Se trata de una tabla de doble entrada, que cruza el comportamiento de un grupo de personas en la prueba criterial que se está diseñando, con el desempeño de ese mismo grupo de personas en la variable que se deberá estimar.

La construcción de esta tabla es relativamente simple, si se retoma el ejemplo de la PM5 y se le quiere utilizar como predictor del rendimiento de 6to grado; lo primero será establecer su *validez funcional,* para lo cual se debe seleccionar a las personas que conformarán el llamado *grupo de criterio*, en quienes se tomarán las medidas en el predictor (PM5) y en la variable a estimar *(rendimiento en 6to grado)*; después de la validación, los resultados se dividirán en categorías; y, finalmente se contará el número (o porcentaje) de personas en cada combinación o celda, tal como se muestra en el cuadro 3.

Cuadro 3
Tabla de expectación para estimar el rendimiento en 6to grado, a partir de la PM5

PM5	Matemáticas de 6to grado		
	Excelente 50-90	Regular 29-499	Deficiente 1-30
20-16	**10** *100*		
15-11	**6** *33*	**12** *67*	
10-6		**5** *25*	**15** *75*
1-5			**15** *100*

Nota: las **negritas son frecuencias** y las *cursivas porcentajes.*

La estimación consiste en reportar el porcentaje de personas en las categorías de interés, permitiendo anticipar el desempeño en matemática de 6to grado de sujetos que aun no ha tomado este curso; así, si un niño obtiene 10 puntos en la PM5, la *tabla de expectación* permite estimar que ese niño tendrá una probabilidad de 25% de obtener una calificación de *regular* en 6to grado; que no existe probabilidad de que califique como *excelente;* y, que hay una probabilidad de 75% de calificar como *deficiente*.

Si bien al agrupar a las personas en una combinación de categorías se sacrifica la precisión que proveen las *calificaciones relacionadas con el contenido* descritas en la sección anterior, con el uso de las calificaciones *relacionadas con el rendimiento* la ganancia consiste en poder anticipar el comportamiento de la persona con solo conocer su desempeño en el instrumento predictor.

b) Nivel predictivo de ejecución

Este representa el segundo tipo de *calificación relacionada con el rendimiento* que se pude preparar. Consiste en indicar para cada posible puntuación generada por el instrumento predictor (PM5), cuál es el promedio de las calificaciones registradas en la variable a predecir (matemática de 6to grado).

Para elaborar una tabla de conversión, de la modalidad *nivel predictivo de ejecución*, se obtienen las calificaciones en el predictor (PM5) y en la variable a predecir (matemática de sexto); luego para cada calificación en el predictor (PM5), supóngase el puntaje 17, se calcula el promedio en matemáticas de 6to grado de los 5 estudiantes que calificaron con 17 en el predictor (55, 57, 57, 58, 59; $\overline{X}$=57,2) y ello se repite para cada uno de los puntajes del predictor. Una vez obtenidos los promedios para cada puntaje predictor, se les puede representar a manera de tubular en un cuadro análogo al 4, o mediante un gráfico como el de la figura 1.

Cuadro 4
Nivel predictivo de ejecución para la PM5

PM5	**6TO GRADO**
20	89,4
19	74,1
18	60,3
17	57,2

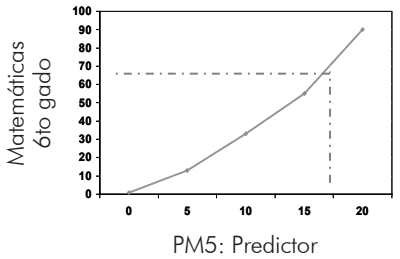

Gráfico 1. Nivel predictivo de ejecución para la PM5

Tanto el cuadro como el gráfico permiten hacer estimaciones. Así, por ejemplo, si se examina a un estudiante que obtuvo un puntaje de 18 en el predictor (PM5*)* sin que haya cursado matemáticas de sexto grado, se puede predecir que obtendrá una calificación de 60,3 puntos de promedio cuando tome ese curso.

Independientemente de que las calificaciones de las *pruebas criteriales* se expresen como *calificaciones relacionadas con el contenido o con el rendimiento*, existen algunas consideraciones que se deben tener en cuenta a la hora de interpretarlas, de modo de evitar incurrir en valoraciones inexactas.

Por un lado, está el hecho de que las pruebas no son perfectamente confiables, lo cual hace indispensable tomar en consideración *los errores de medición* a la hora de estimar la calificación de un examinado. Los errores en que se puede incurrir debido a la inconfiabilidad de las pruebas, se subsanan al presentar los cuadros de conversión que utilizan el sistema de bandas en lugar de calificaciones exactas. Por otro lado, está la tentación de generalizar los resultados de la prueba a otras situaciones, sin haber realizado los estudios de validez pertinentes, pero mientras se carezca de tales datos solo se debe emplear las calificaciones en el test, analizándolas en función de ese dominio y con ningún otro fin o variable.

V. Etapa de revisión

Cuando los resultados informan que el instrumento es *válido* y *confiable,* se han preparado los datos para su interpretación y se está en la posibilidad de usarlo para los fines para los cuales se diseñó. Es entonces cuando se

procede a entregarlo a una casa editorial o revista especializada. Hasta ese momento, el instrumento ha estado exclusivamente en manos del diseñador; no obstante, en lo sucesivo entran en juego los usuarios o examinadores y queda bajo la responsabilidad de todos observar el óptimo funcionamiento del recurso de medición. Si la impresión inicial era que el instrumento quedaba listo y que así sería por tiempo indefinido, la misma naturaleza de las pruebas indicará que ese recurso queda ahora bajo la luz de una variedad de actores quienes deberán "revisar" su desempeño.

Como puede apreciarse, esta etapa se diferencia de las cuatro anteriores en que aquellas son circunstanciales y participa exclusivamente el equipo que elaborado la prueba, mientras que la *etapa de revisiones* es permanente y contribuyen, como se dijo, además de los autores, los usuarios del instrumento. En tal sentido, debe rescatarse que los usuarios y los autores de la prueba mantienen el compromiso ético de informarse mutuamente los cambios observados en el comportamiento del instrumento, de modo de mantenerlo con propiedades psicométricas adecuadas y actualizado.

Cuando se observan inadecuaciones en el desempeño de un instrumento en *etapa de revisiones,* hay que determinar el nivel de la falla a los fines de introducir los estudios y ajustes del caso: si las observaciones son leves se realizarán ajustes ligeros, pero si son sustanciales es probable que haya que realizar correcciones al nivel de la estructura de los ítems. Con lo anterior se quiere reflejar que el nivel de falla determinará qué tan atrás se deba regresar en las etapas de construcción del instrumento para realizar los ajustes del caso.

Pruebas criteriales *versus* pruebas normativas

En las secciones precedentes de este documento se expresó con detalles cómo se conceptualizan las pruebas referidas a un criterio, las etapas mediante las cuales toma lugar su construcción y el modo como se deben llevar a cabo los estudios de ítems, confiabilidad y validez, con lo cual se pretendió poner sobre el tapete que son instrumentos de medición en su propio derecho y que si bien surgieron como una rama de los tests normativos, siguen caminos de diseño y construcción que los hacen paritarios en términos de la posibilidad de garantizar su calidad.

La presente sección pretende rescatar las diferencias y similitudes entre los dos tipos de instrumentos, comparando aspectos relativos sus objetivos, interpretación de los números que generan, características de la selección de

las muestras de trabajo y énfasis en los procesos de confiabilidad y validez, como un modo de terminar de deslastrar a un tipo de prueba de la otra.

La diferenciación entre pruebas criteriales y normativas no puede hacerse recurriendo simplemente a la inspección de sus protocolos de preguntas. Se debe, en cambio, precisar los propósitos para los cuales fueron diseñadas, la información que proveen y la manera como fueron construidas y analizadas psicométricamente.

Los tests criteriales se utilizan cuando el objetivo es describir lo que una persona puede o no hacer respecto a un dominio o cuando se necesita clasificar a las personas en atención a las conductas de su repertorio; de modo que informan, por ejemplo, que un niño exhibe unos patrones conductuales que apuntan a un trastorno por déficit de atención con hiperactividad, o con predominio hiperactivo-impulsivo (HI). Por su parte, las pruebas normativas se emplean cuando el interés es identificar diferencias individuales a partir de la comparación del comportamiento entre las personas. Es así como se puede decir, por ejemplo, que un niño en la variable razonamiento numérico supera o igual al 90% de los sujetos de su grupo normativo.

Como se pudo notar antes, los tests referidos a un criterio comparan la ejecución de una persona con estándares establecidos a partir del dominio examinado, por lo que son independientes del grupo de sujetos; dichos estándares los especifica el constructor en las etapas iniciales de elaboración del test. Las pruebas normativas, en cambio, usan estándares que se establecen en función del comportamiento del grupo después de administrarlas, para poder producir esos estándares de comparación; en suma, es el grupo quien los provee (Popham, 1978).

Las pruebas criteriales exigen una evaluación exhaustiva del dominio que examinan, de modo de disponer de un muestreo representativo de las tareas o conductas cubiertas por la prueba; así, lo que sea cierto para esa muestra de conductas también lo será para el universo de conductas. Las pruebas normativas atienden menos al contenido por sí mismo y se concentran en demostrar numéricamente que existe una relación entre cada elemento de la prueba con la variable examinada.

Las pruebas normativas exigen disponer de una muestra representativa de las personas, por lo que recurren a procedimientos probabilísticos a la hora de obtener los datos para preparar los análisis de ítems, las normas y las

propiedades psicométricas. Las pruebas criteriales, por su parte, emplean muestreos intencionales de las personas, de manera de conformar grupos de criterio (Glaser y Nitko, 1971).

Los reactivos de las pruebas criteriales se ajustan hasta que arrojen niveles de dificultad extremos; así, son deseables aquellos que tengan un nivel de dificultad de 1, para un grupo que posea el dominio examinado y de 0 para el que no lo posea; en la medida en que los reactivos se alejen de tales valores extremos de dificultad, serán eliminados. En las pruebas normativas, por el contrario, los planteamientos con niveles de dificultad extremos son completamente descartables y se busca conformar la prueba con reactivos que arrojen una dificultad promedio ($p = 0,50$), ya que ellos serán los que produzcan mayor varianza y permitirán una mejor diferenciación entre las personas (Carver, 1974).

Las pruebas criteriales y normativas son similares en el sentido de que ambas conllevan estudios de validez con el objetivo de demostrar que cumplen los objetivos para los cuales fueron diseñadas. En el caso de las pruebas criteriales, se alude a una validez descriptiva, funcional y de dominio; y, en el caso de las pruebas normativas, se hace referencia a validez de contenido, predictiva y de constructo. El tipo de estudio realizado depende de los objetivos de la prueba (Carver, 1974).

Es posible identificar ciertas analogías en los objetivos de los estudios de validez para las pruebas criteriales y normativas. Así, el propósito de la validez descriptiva (criterial) y la de contenido (normativa) es procurar que la prueba cubra apropiadamente el contenido examinado. El objetivo de la validez funcional (prueba criterial) y predictiva (prueba normativa) es proveer al instrumento de información en torno a la precisión de las predicciones que se pueden hacer con él. La misión de la validez de dominio (prueba criterial) y de constructo (prueba normativa) es informar que se mide un dominio válido en términos teóricos (Popham, 1978).

Con los estudios de confiabilidad de las pruebas criteriales y normativas se busca exactamente el mismo objetivo: determinar la consistencia de las calificaciones, haciendo el énfasis correspondiente. De este modo, si el interés es precisar si el muestreo de tiempo afectará la consistencia, debe emplearse un procedimiento de retest, pero si lo que interesa es conocer qué tanto afecta a la consistencia el muestreo de contenido, resultará pertinente aplicar un procedimiento de pruebas paralelas.

Las unidades en las que se expresan las pruebas criteriales están apegadas a las consideraciones de su constructor, quienes recurren a las calificaciones de pase y los porcentajes de respuestas correctas cuando tienen fines diagnóstico; y a las tablas de expectación y los niveles predictivos de ejecución, cuando tienen fines predictivos. Por su parte, las pruebas normativas usan escalas muy conocidas dentro de la comunidad científica; entre ellas se tienen las calificaciones estándar y las escalas de desarrollo. Para efectuar predicciones recurren a la ecuación de regresión simple o múltiple e, incluso, calificaciones de corte (Lezama, 2011).

Se debe concluir señalando que si bien las pruebas criteriales y normativas persiguen objetivos distintos que determinan los pasos comprometidos en su construcción, ambas exigen propiedades psicométricas de validez y confiabilidad, que pueden ser cumplidas con total rigurosidad científica.

Referencias

Almerich, G. y Bo Bonet, R. (2006). Efecto de la forma de la distribución y de la media en el índice Po de Huynh**.** *Relieve, 12*, 151-166. http://www.uv.es/RELIEVE/v12n1/RELIEVEv12n1_6.htm.

Backoff, E., Sánchez, A., Peón, M., Monroy, L. y Tanamachi, M. (2006). Diseño y desarrollo de los exámenes de la calidad y logro educativos. *Revista Mexicana de Investigación Educativa, 11,* 617-638.

Berk, R. (1980). A framework for methodological advances in criterion-referenced testing. *Applied Psychological Measurement*. 4, 563-573.

Carver, R. (1974). Two dimensions of tests: Psychometic and edumetric. *American Psychologist, 29,* 512-518.

Coccaro, E., Kavoussi R., Berman M.E, Lish J. (1998). Intermittent explosive disorder-revised: Development, reliability, and validity of research criteria. Comtemporary Psychiatry, 39, 368-76.

Conners, C. (1989). *Conner´s ratign scales*. Toronto, Ontario: Multi-Health Systems.

Dembo, M. (1983). Dos estrategias de estrategias de medición en psicología. *Psicología, 10,* 41-57.

Glaser, R. y Nitko, A.J. (1971). Measurement in learning and instruction. En R.L.Thorndike (Ed.). *Educational measurement.* Washington, D.C.: American Council on Education.

Lacasella, R. (2000), *Metodología para el estudio del desarrollo infantil desde la perspectiva conductual*. Caracas: Fondo Editorial de la Facultad de Humanidades y Educación

Lezama, L. (2011). Puntuaciones relacionadas con las normas. *Psicología, 30,* 107-143.

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

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