La Fiabilidad y la validez desde la perspectiva criterial

Juan MATEO ANDRES

Universidad de Barcelona.

FIABILIDAD

CONCEPTO

Dentro del modelo psicométrico clásico se concibe la fiabilidad en términos de los componentes de la varianza relativos a las puntuaciones observadas, verdadera y de error, y se expresa a nivel de cálculo como confiabilidad, equivalencia, consistencia, precisión o estabilidad, como diversas acepciones respecto a si la prueba mide siempre en el mismo sentido y/o si sus componentes miden el mismo aspecto.Esta concepción genérica de fiabilidad no se ha trasladado en su totalidad a los modelos criteriales, donde el interés fundamental no es establecer comparaciones entre sujetos ni tampoco diferencias entre individuos. Lo que importa en este contexto es buscar el nivel de eficacia individual, si dominan o no una habilidad o qué es lo que no dominan, se sugirere incluso en este contexto el utilizar la denominación de índices de *acuerdo* en lugar de índices de fiabilidad.

Dada esta finalidad, los instrumentos de referencia criterial y para poder determinar un tipo de dominio, se construyen de tal forma que se produce una reducción en la variabilidad de las puntuaciones de los sujetos (Se utilizan mayor número de ítems por objetivo, abarcan contenidos más reducidos, no se seleccionan los ítems por su nivel de dificultad o su poder descriminativo, etc), y conocido es que el coeficiente de correlación depende de la variabilidad de las puntuaciones y si ésta es baja, también lo serán los correspondientes coeficientes de fiabilidad obtenidos a partir de ellas. Todo ello ha llevado a decidir (Popham y Husek 1969), que los métodos habituales (el correlacional por ejemplo) no son los más adecuados para fiabilizar estos instrumentos, debiéndose por tanto el explorar nuevas formas.

Hambleton, Swaminatham, Algina y Coulson (1978, p. 15-23) distinguen tres conceptualizaciones de fiabilidad en torno a los tests referidos al criterio:

1.- FIABILIDAD COMO MEDIDA DE ACUERDO EN LAS DECISIONES DE CLASIFICACION

En esta etapa se enfatiza la consistencia de las decisiones a partir de las cuales se divide a los individuos en "dominadores - no dominadores", cuando se repite un mismo instrumento a un mismo grupo, ya sea a través de formas paralelas o aleatoriamente paralelas.

Estos índices se denominan "índices de pèrdida de umbral" e implican:

a) Establecimiento de puntuaciones de corte (cut-off), en función de la cual se tomarán las decisiones y por tanto una clasificación dicotómica.

b) Las pérdidas asociadas a los errores de decisión se consideran de igual importancia.

Este concepto de fiabilidad fue propuesto por vez primera por Hambleton y Novick (1973 p. 168).

2.- FIABILIDAD DE LAS PUNTUACIONES "CUT-OFF"

En este enfoque se enfatiza la consistencia de las desviaciones de las puntuaciones respecto a las puntuaciones cut-off, a través de formas paralelas o aleatoriamente paralelas.

Estos índices se denominan "índices de pérdida de varianza" e implican:

a) Una puntuación de corte.

b) Las pérdidas asociadas a los errores de decisión no se consideran de igual importancia.

3.- FIABILIDAD DE DOMINIO DEL INDIVIDUO

Se enfatiza la consistencia de las puntuaciones respecto al dominio de los individuos, estimada a partir de las puntuaciones de cada estudiante en particular, a través de las dos formas de la prueba y sin hacer referencia alguna a las puntuaciones de corte.

Podríamos resumir lo anterior mediante el I cuadro

B.- PROCESO

Según Berk (1980) el proceso de fiabilización de un instrumento de referencia criterial pasa por dos etapas:

1.- Se escoge la acepción apropiada de fiabilidad, según se desee enfatizar en una de las tres conceptualizaciones de fiabilidad antes mencionadas.

2.- Se elige el índice estadístico específico dentro de cada acepción.

CUADRO I

CONCEPTUALIZACIONES DE LA FIABILIDAD EN LOS TESTS REFERIDOS AL CRITERIO

Categoría / Carácter Carácter/Categoría	Pérdida de umbral (Threshold Loss)	Pérdida Varianza (Squard-error Loss)	Estimación de la puntuación de dominio (domain Score estma.)
Interpretación de la puntuación.	Cada puntuación individual se refiere a la puntuación de corte.	Cada puntuación individual se refiere a la puntuación de corte.	Cada puntuación se utiliza para estimar la puntuación de dominio.
Tipo de decisión: Información que requiere.	Clasificación como "domina" "No domina"	Acuerdo de la clasificación a partir de la puntuac. total.	Nivel de competencia en los contenidos de dominio.
Pérdidas asociadas con los errores de decisión.	Las pérdidas son igualmente importantes, sin importar la cantidad.	Las pérdidas relativas a los estudiantes mal clasificados que están muy por encima o por debajo de la punt. de corte son más importantes que las debidas a otros estudiantes mal clasificados.	Consideración de pérdidas diversas a partir del procedimiento de estimación.

C.- INDICES DE FIABILIDAD

Berk (1980) en su artículo "A consumers guide to criterion referenced test reliability" indica como entre los años 70 y 80 surgen más de una docena de estadísticos diferentes para la estimación de la fiabilidad de los tests referidos al criterio, que podríamos sintetizar de la siguiente forma:

Subkoviak (1982), al que Berk considera el mejor representante de la primera acepción, indica como el significado de la fiabilidad como medida de acuerdo en las decisiones que se toman a partir del test, se refieren a la consistencia con que los individuos son clasificados como que "dominan" o que "no dominan" en los objetivos medidos en la prueba. Autores como: Carver (1970, 1974), Hambleton y Novick (1983), Swaminatham, Hambleton y Algina (1974), Popham (1978, 1980), Marshall y Haektel (1976), Huynk (1976) y Subkoviak (1976) han desarrollado diferentes procedimientos de cálculo para expresar este tipo de fiabilidad.

Dentro de la segunda conceptualización de fiabilidad, encuadraríamos a Livingston (1972), basados en la teoría de la Generalizabilidad estarían los desarrollados por Brennan y Kane (1977) y Kane y Brennan (1980) y los métodos propuestos por Wine (1971) y Lovet (1977). Siendo Brennan, posiblemente, el mejor representante de este tipo de conceptualización.

Finalmente, en la tercera conceptualización de fiabilidad entendido como fiabilidad de dominio, estarían los desarrollados por: Cochran (1963), Millman (1974), Lord (1955, 1957, 1959), Lord y Novick (1968), Brennan (1980).

Resumimos lo anterior mediante el siguiente cuadro:

INDICES DE FIABILIDAD DE LOS TESTS REFERIDOS AL CRITERIO

Consistencia de las Decisiones	- Carver (1970) - Hambleton y Novick (1973) $\hat{p}o$ - Swaminatham, Hambleton y Algina (1974) $\hat{k}$ (basada en Cohen, 1960) - Marshall (1976) ß ($\hat{p}o$) - Subkoviack (1976, 1980) $\hat{p}o$, $\hat{k}$ - Huynh (1976) $\hat{p}o$ / $\hat{k}$- Huynh (1977) KM
Consistencia de las Desviaciones	- Van der Linder ans Mellenbergh (1978) L (ζ, x) - Livingston (1972) K^2 (x,t) - Lovett (1977,1978) - Brennan (1977, 1978, 1980) Φ (λ) ó - Brennan and Kane (1977 a 1977b) ID (γ) ó M (c) - Kane and Brennan (1980)
Consistencia de las estimaciones de dominio (errores estandard)	Individual:Berk (1980) Cochram (1963) Millman (1974) $\hat{\varepsilon}p$ Lord (1955-1957-1959) $S.E._M$ (Xa) Grupal: Lord y Novick (1968) $\hat{o}_E$ Brennan (1980) $\hat{o}_{(\Delta)}$ y Φ

D.- DESARROLLO DE ALGUNOS INDICES DE FIABILIDAD

METODO DE SUBKOVIAK (índice de fiabilidad como medida de acuerdo en las decisiones de clasificación)

Requiere las puntuaciones de un solo test. Los pasos son los siguientes:

1.- A partir de las puntuaciones obtenidos por el grupo de sujetos y de las frecuencias de cada puntuación se calcula la $\hat{\mu}$ y el coeficiente de Kuder Richardson nº 20.

2.- Los ítems se consideran como una muestra de un actual o hipotético universo de ítems semejantes. A partir de los valores obtenidos en el primer punto se calcula $\hat{p}x$. Este estadístico es una estimación de la proporción de ítems en el universo, que puede esperarse que una persona con una puntuación x dada conteste correctamente.

Así, $\hat{p}x$ puede considerarse como la probabilidad de contestar correctamente el ítem.

Px = KR 20 (x/n) + (1 - KR 20) (μ/n)

X = puntuación

n = nº de ítems

3.- Si la probabilidad de que un estudiante responda correctamente un ítem es $\hat{p}x$ (calculado en el punto 2), ¿Cuál es la probabilidad de que el estudiante pueda responder correctamente un número de ítems igual o mayor que la puntuación cut-off, y, por tanto, sea clasificado como "dominados"? Se trata pues de calcular esta probabilidad mediante la distribución binominal, siendo n el nº total de ítems, $\hat{p}x$ la probabilidad de acertar un item, y $\hat{p}x \geq$ cut-off. Así, se obtendrá la distribución de $\hat{p}x$ o probabilidad de ser consistentemente clasificado como master. Esta será

$$Pi\ (x \geq c) = \sum_{x=c}^{n} \binom{n}{x} \hat{\zeta}^n (1 - \hat{\zeta})^{n-x}$$

siendo

Pi ($x \geq c$) = probabilidad que tiene un individuo de obtener la puntuación igual o superior a la puntuación de corte (c)

$\hat{\zeta}$ = puntuación verdadera

x = puntuación observada

n = número de ítems del test

4.- La probabilidad de ser consistentemente clasificado en dos tests independientes es $\hat{p}^2_x$, inversamente, la probabilidad de que un estudiante sea consistentemente clasificado como "no dominados" será: $(1- \hat{p}_x)$

La probabilidad de ser consistentemente clasificado para el sujeto será:

$$\hat{p}^2_x + (1 - p_x)^2 : 1 - 2 (\hat{p}_x - \hat{p}^2_x)$$

5.- La probabilidad de consistencia en las clasificaciones para todo el grupo de sujetos será $\hat{p}_o$, y se obtendrá:

$$\hat{p}_o = \frac{\sum N_x - 1 - 1 (\hat{p}_x - \hat{p}^2_x)}{N}$$

siendo N = nº de sujetos

6.- Se obtendrá $\hat{p}_c$ probabilidad de clasificación consistente debida al azar

$$\hat{p}_c = 1 - 2 \frac{\frac{\Sigma N_x \; p_x}{N}}{\frac{\Sigma N_x - p^2 x}{N}}$$

7.- Finalmente se obtendrá $\hat{K}$ (coeficiente general de acuerdo)

$$\hat{K} = \hat{p}_o - \hat{p}_c / 1 - \hat{p}_c$$

- INDICE DE BRENNAN Y KANE (Fiabilidad de las puntuaciones cut-off)

$$\phi = \frac{\sigma^2(p) + (\mu - 1)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\Delta)}$$

siendo:

$\sigma^2(p)$ = varianza personas
μ = media total
λ = puntuación cut-off

$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(p - I)$ = varianza de ítems más varianza de personas por ítems.

- INDICE DE BRENAN, COCHRAN Y MILLMAN (fiabilidad entendida como fiabilidad de dominio del individuo)

$$\hat{\varepsilon}_p = \sqrt{\frac{p\,q}{n - 1}}$$

siendo:
p = porcentaje de ítems contrastados
q = 1 - p
n = nº de ítems.

VALIDEZ

CONCEPTO

La validez es sin duda el aspecto o característica más relevante de un instrumento. Hace referencia y aparece vinculado al concepto de "verdad o veracidad" de "adecuación" de "utilidad" y de "servicio".

La determinación de la validez de un instrumento es un proceso continuo, que abarca, incluye y debe formar parte y estar presente a lo largo de todo el proceso de elaboración, puesta en acción y utilización del mismo. Implicará estudios empíricos y lógicos y es una cuestión de grado.

La validez de un instrumento indica la adecuación del mismo para medir lo que se pretende y para la finalidad perseguida. Lo que realmente se valida, sin embargo, no es el instrumento en si, sino, como señal Cronbach (1971) *la interpretación de los datos obtenidos por medio de un procedimiento especificado.*

EVIDENCIAS DE VALIDEZ

Hasta muy recientemente se hablaba de los diferentes *tipos de validez*, sin embargo desde el año 1985 la Joint Technical Standards for Educational and Psychological Testing de la American Psychological Association, se refiere a la validez como un *concepto unitario* y destierra el concepto de tipos diferentes de validez, sustituyéndolo por el de aproximaciones o evidencias distintas que nos conducen a ella.

Las tres aproximaciones básicas para acumular evidencias que den soporte a la validez de una interpretación son las de:

A) CRITERIO

B) CONTENIDO

C)CONSTRUCTO

De las tres evidencias la de contenido es sin duda la fundamental para los tests de referencia criterial, será por ello que le dedicaremos atención específica. Dentro del marco de los Tests referidos al criterio, algunos autores, como Popham emplean el término de evidencia descriptiva, en lugar de contenido, ya que (1983), "esta evidencia de validez trata de comprobar hasta qué punto un test basado en criterios mide realmente lo que su esquema descriptivo dice que mide"

Dado que el objetivo de este tipo de instrumentos es proporcionar una estimación de las capacidades o conductas que cada sujeto posee respecto a un determinado dominio o universo definido, el muestreo de las mismas será una cuestión fundamental que habrá que plantearse, en especial su representatividad respecto del universo de conductas.

Habrá que dejar bien sentado:

— Cúal es el universo de dominio, y qué elementos y especificaciones comporta, como se agrupan, etc.

— Cómo se extrae la muestra.

—Qué tipo de inferencias, permite la muestra de acuerdo con su estructura, agrupación o cluster de ítems, etc.

—Qué significan las respuestas, etc.

La evidencia descriptiva se orientará fundamentalmente hacia: *la relevancia del mismo* y a su *representatividad.*

(Livingston, 1977, Guion, 1977)

PROCESO

La evidencia descriptiva o de contenido en los tests referidos al criterio implicará determinar:

1.- La especificación del dominio.

2.- La validez de los ítems.

3.- La calidad técnica de los ítems.

4.- La representatividad de los ítems del test.

1.- Especificación del dominio

Siguiendo a Linn (1980) habrá que crear una lista exhaustiva de ítems o reglas satisfactorias de generación de los mismos para después a partir de la teoría de la probabilidad hacer inferencias a partir de las puntuaciones.

Esta posición contrasta con la tradicional que consideraba el conjunto de ítems del test, como una muestra extraída de un dominio que podía considerarse infinito.

2.- Validez de los ítems.

Una vez se ha descrito el universo, dominio, o campo, mediante el posible empleo de especificaciones de dominio, generación de las formas de ítems, objetivos ampliados, etc, se elaboran los ítems. El objetivo fundamental será establecer la *congruencia entre el ítem y el objetivo o especificaciones de dominio.* En la determinación de esta congruencia se emplearán tres criterios: la conducta, el contenido y la clasificación jerárquica.

Hambleton (1982) señala dos tipos de procedimientos:

A) Procedimientos judiciales ("a priori")

B) Procedimientos empíricos ("a posteriori")

3.- Calidad técnica de los ítems.

Se trataría de la revisión técnica de los ítems sobre su grado de adecuación. Hay muchas variables que influyen en *esa adecuación.* La lecturabilidad, el formato de los mismos, el tiempo de ejecución, el tipo de distractores que se incluyen, la presentación, etc. Todas estas son variables que habrá que tener presentes, tanto cuando se validan mediante procedimientos judiciales, como cuando se validan mediante procedimientos empíricos.

4.- Representatividad de los ítems.

Una vez especificado el dominio, habrá que juzgar si la muestra de ítems es representativa del mismo, para ello deberemos recurrir a sistemas derivados de la lógica de la probabilidad muestral.

Cronbach (1971 pp. 456) propone un sistema muy ingenioso, consistente en construir de forma independiente dos test en donde se parta de la misma definición de contenido relevante, se utilicen las mismas reglas de muestreo, instrucciones a las que revisan los ítemas y especificaciones para juzgar e interpretar los datos. Si la especificación de dominio es clara y si la muestra es representativa, los dos tests

deberían ser equivalentes, y puede comprobarse ambos tests a un mismo grupo de examinados y comparando los dos grupos de puntuaciones del test.

Finalizamos nuestra reflexión indicando que a pesar de que hemos enfatizado la evidencia de contenido o descriptiva en el contexto de los tests referidos al criterio, deberan utilizarse todas las demás para aproximarse a ese concepto unitario de validez a que hacíamos referencia al principio. Suscribiremos así las palabras de Linn (1980, p. 559) cuando nos señala: "La validez de contenido proporciona una excelente fundamentación para los tests referidos al criterio, pero para soportar la validez de un instrumento, se necesita, además, la validez de las inferencias y de los usos del mismo".

BIBLIOGRAFIA

BERK, R. A. (Ed) (1980) *Criterion Referenced Measurement: the State of the art.* Baltimore: Johns Hopkins Univ. Press.

BRENNAN, R. L. (1980) Aplications of generalizability theory en R. A. Berk (ed) *Criterion referenced measurement: the state of the art.* Baltimore: Johns Hopkins. Univ. Press.

BRENNAN, R. L. y KANE, M. T. (1977) "An index of dependability for mastery tests" *Journal of Educational Measurement,* 14, 277-87.

CARVER, R. P. (1970) "Special problems in measuring change with psychometric devices". En *Evaluative research : Strategies and methods.* American Institutes for Research, 48-63. (1974) "Two dimensions of tests" "Psychometric and Edumetric" *American Psychologist,* 29, 512-518.

COCHRAN, W. G. (1963) *Samplling techniques* New York : Wiley

CRONBACH, L. J. (1971) "Test validation" en R. L. Thorndike (Ed) *Educational Measurement* Washington: American Council on Education.

GUION, R. M. (1977) "Content validity : The source of my discontent" *Applied Psychological Measurement,* 1, pp. 10.

HAMBLETON, R. K. (1982) "Advances in criterion referenced testing technology" en C. R. Reynolds y T. B. Gutkin (Eds.) *The handbook of school psychology.* New York , Wiley p. 351-379.

HAMBLETON, R. K. y NOVICH, M. R. (1973) "Toward an integration of theory and method for criteriorn-referenced test" *Journal of Educational Measurement,* .10, 159-70.

HAMBLETON, R. K., SWAMINATHAN, H., ALGINA, J. y COULSON, D.B. (1978) "Criterion-referenced testing and measurement : A review of Technical Issues and developments" *Review of Educational Research,* VOL 48, 1, P. 1-47.

HUYNH, H. (1976) "On the reliability of decisions in domain referenced testing" *journal of Educational Measurement,* 13, p. 253-64.

KANE, M. T. y BRENNAN, R. L. (1980) "Agreement coefficients as indices of dependability for domain-referenced tests". *Applied Psychological Measurement,* vol. 4, 1, 105-26.

LINN, R. L. (1980) "Issues of validity for criterion-referenced measures" *Applied Psychological Measurement,* 4, 575-581

LIVINGSTON, S. A. (1972) "Criterion-referenced applications of classical test theory". *Journal of educational Measurement,* 9, p. 13-26.

(1977) "Psychometric techniques for criterion-referenced testing and behavioral assessment" en J. D. Cone y R. P. Hawkins (Eds). *Behavioral assessment: New directions in clinical psychology.* New york: Brunner-Magel, 308-329.

LORD, F. M. (1955) "Sampling fluctuations resulting from the sample of test items". *Psychometrika,* 20, 1-22. (1975)" Do tests of the same length have the same standard error of measurement? "*Educational and Psycholigical Measurement,* 17, 510-21. (1959) "Tests of the same length have the same standard error of measurement? "*Educational and Psychological Measurement,* 19, 233-239.

LORD, F. M. y NOVICK, M. R. (1968) *Stadistical theories of Mental Test scores.*Reading, µassachusetts: Eddison-Wesley.

MARSHALL, J. L. y HAERTEL, E. H. (1976) *The mean split half coefficient of agreement: A single administration index of reliability for mastery test.* Manuscrit. University of Wisconsin.

MILLMAN, J. (1974) "Criterion-referenced measurement" en W.J. Popham (Ed.) *Evaluation: Current Applications.* Berkeley (CA), Mc Cutchan Publishing Company

POPHAM, W. J. (1978) *Criterion-referenced Measurement.* Englewood Cliffs, (M. J.) prentice-Hall. (1980) "Content domain specifications" En R. A. berk (Ed.) *Criterion-referenced measurement: The state of the art.* Baltimore (M. D.) Johns Hopkins University Press, p. 15-31. (1983) *Evaluación basada en criterios.* Madrid: Magisterio Español.

POPHAM, W. J. y HUSEK, T. R. (1969) "Implications of criterion-referenced measurement" *Journal of Educational Measurement,* 6-1, p. 1-9.

SUBKOVIAK, M. J. (1976) "Estimating reliability from a single administration of a mastery test" *Journal of Educational Measurement,* 13, p. 265-76.