Fiabilidad

Maite Barrios
Antoni Cosculluela

PID_00198628

Universitat Oberta de Catalunya

Índice

Introducción

En el lenguaje cotidiano el término *fiabilidad* se asocia a algo que funciona de manera correcta. Nos fiamos de nuestro despertador si suena a la hora que se ha programado, de la báscula si nos proporciona sin error nuestro peso, incluso consideramos que contamos con un buen amigo si siempre nos apoya cuando lo necesitamos. Si el despertador, la báscula y nuestro amigo no se comportan de la manera "correcta", consideramos que no son fiables y en consecuencia decidimos que no podemos confiar en ellos.

En psicometría nos referimos a la fiabilidad como aquella propiedad que valora la consistencia y precisión de la medida. En consecuencia, si la medida toma valores consistentes y precisos, creemos que podemos confiar en los resultados obtenidos cuando se aplica un test. No obstante, sabemos que cualquier proceso de medida (se esté midiendo un objeto físico o un aspecto psicológico) se asocia a algún grado de error. La medida perfecta no existe. El estudio de la fiabilidad de un instrumento de medida debe permitir conocer hasta qué punto los resultados que se obtienen a partir de su aplicación están afectados por el error que se ha cometido al medir. Si el error es pequeño, podemos confiar en el resultado del test; si el error es grande, el proceso de medición deja de tener sentido. En este capítulo se trata el tema de la fiabilidad desde dos vertientes: por un lado, se presenta la fiabilidad desde la perspectiva de la teoría clásica de los tests (TCT), centrándonos en los test referidos a la norma (TRN), para, por otro lado, abordar el tema de la fiabilidad según los tests referidos al criterio (TRC).

Desde la perspectiva de la TCT se presenta lo que se entiende por error de medida y los diferentes tipos de error de medida que se pueden cometer al aplicar un test. A continuación, se describe el modelo lineal propuesto por Spearman y cómo a partir de él se deriva el coeficiente de fiabilidad. Nos detendremos en cómo interpretar un coeficiente de fiabilidad, así como en las diferentes estrategias que se han ido desarrollando para calcularlo: test-retest, formas paralelas y consistencia interna. A continuación, tratamos tres de los factores que influyen en la fiabilidad (variabilidad de las puntuaciones obtenidas en el test, la longitud del test y las características de los ítems que lo componen). Para acabar con la TCT, se presentan dos procedimientos para valorar la puntuación verdadera de un sujeto: la estimación que asume la distribución normal del error aleatorio y la estimación a partir del modelo de regresión lineal.

Veremos una manera diferente de abordar la fiabilidad cuando los tests que se emplean son instrumentos cuyo objetivo es valorar la competencia de las personas en algún dominio concreto de conocimiento, los denominados TRC. Para contextualizar la fiabilidad en los TRC, en primer lugar, nos detenemos en los conceptos básicos que caracterizan este tipo de tests y, en segundo lugar,

describimos los tres procedimientos más clásicos para abordar su fiabilidad: aquellos procedimientos que requieren dos aplicaciones del test para valorar la consistencia de la clasificación, aquellos que solo requieren una única aplicación y aquellos en los que entra en juego el papel de los evaluadores. En el último apartado del módulo, se describen los métodos que más frecuentemente se utilizan para determinar el punto de corte que permite una mejor clasificación entre aquellos individuos que son competentes en el criterio de interés y aquellos que no lo son.

1. Concepto de fiabilidad según la teoría clásica

Según la teoría clásica de los tests, la fiabilidad de un test está relacionada con los errores de medida aleatorios presentes en las puntuaciones obtenidas a partir de su aplicación. Así, un test será más fiable cuantos menos errores de medida contengan las puntuaciones obtenidas por los sujetos a quienes se les aplica. Dicho de otro modo, la fiabilidad de un test será su capacidad para realizar medidas libres de errores.

1.1. El error de medida

Todo instrumento de medida debe garantizar, con más o menos rigor, que las medidas que obtenemos con su aplicación se corresponden con el verdadero nivel o valor de la característica evaluada. Así, si queremos medir la temperatura del agua del mar un día de un caluroso mes de agosto, necesitaremos un termómetro que nos permita obtener este dato. Si lo hacemos con el termómetro que compramos en unos grandes almacenes para medir la temperatura del agua de la bañera de casa, seguramente obtendremos un valor que será menos preciso que si lo hacemos con el termómetro que utiliza el servicio de meteorología para tomar estas medidas. En cualquier caso, seguramente tanto uno como otro termómetro medirán con un cierto grado de imprecisión, posiblemente más elevado en el primer caso que en el segundo, pero ninguno exento de una cierta desviación respecto a la verdadera temperatura del agua. Si la medida la hiciéramos utilizando un sofisticado instrumental cedido por la NASA, seguramente tendríamos bastantes más garantías de que la temperatura obtenida se corresponde con mucha más precisión con la verdadera.

Por lo tanto, cualquier proceso de medida de una característica de los objetos o de los sujetos lleva inherente un cierto error en su medición. Podemos encontrar instrumentos de medida con más o menos capacidad para minimizar estos errores, pero difícilmente podremos encontrar uno que los elimine del todo.

En nuestro ámbito de la psicología, donde las variables que medimos habitualmente son características propias de los sujetos, relacionadas con sus rasgos de personalidad, sus capacidades cognitivas, sus estados de ánimo, etc., y donde los instrumentos utilizados para la medición son generalmente los tests, aún resulta más evidente que las medidas que hacemos de estos atributos estarán también afectadas por ciertos errores. Esto provocará que las puntuaciones obtenidas con las administraciones de estos tests no se correspondan exactamente con los verdaderos niveles de los sujetos en la característica medida.

En cualquier caso, algunos de estos errores propios de toda medición pueden responder a factores sistemáticos que tendrán una posible causa en el propio proceso de medida, en el instrumento utilizado o en ciertas características de

los objetos o sujetos medidos. Así, si el termómetro con el que medíamos la temperatura del agua del mar tiene un error de construcción que hace que siempre mida un grado más del real, este error afectará por igual a toda medición realizada con él, y se podrá eliminar haciendo una buena calibración del aparato. Otros errores no tienen este componente sistemático, sino que son aleatorios, indeterminados y no responden a ningún factor que pueda ser conocido, y por lo tanto eliminado. Estos errores aleatorios son los que están implicados en el concepto de fiabilidad.

1.2. El coeficiente de fiabilidad y su interpretación

Desde la teoría clásica de los tests (TCT) de Spearman, se define el coeficiente de fiabilidad de un test $\rho_{xx'}$ como la correlación entre las puntuaciones obtenidas por un grupo de sujetos en dos formas paralelas del test.

Según la definición de formas paralelas de un test de la TCT, si un test tuviera una fiabilidad perfecta, las puntuaciones obtenidas por un sujeto en las dos formas paralelas del test deberían ser idénticas, y por lo tanto la correlación entre las puntuaciones de un grupo de sujetos en estas dos formas paralelas del test sería 1 ($\rho_{xx'} = 1$). Cualquier valor inferior a 1 se deberá a los errores aleatorios propios del instrumento de medida.

A partir de la definición anterior del coeficiente de fiabilidad, y teniendo en cuenta los supuestos de la TCT, también podemos expresar el coeficiente de fiabilidad como el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas. En este sentido, el coeficiente de fiabilidad se puede interpretar como la proporción de varianza de las puntuaciones verdaderas (σ_v^2) que hay en la varianza de las puntuaciones empíricas (σ_x^2):

$$\rho_{xx'} = \frac{\sigma_v^2}{\sigma_x^2}$$

De la expresión anterior y de las consecuencias derivadas de la TCT podemos deducir fácilmente que este coeficiente de fiabilidad será igual a 1 menos la proporción de la varianza de los errores (σ_e^2) que hay en la varianza de las puntuaciones empíricas.

$$\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

Un índice directamente relacionado con el coeficiente de fiabilidad es el índice de fiabilidad (ρ_{xv}) que se define como la correlación entre las puntuaciones empíricas de un test y las puntuaciones verdaderas. Este índice de fiabilidad es igual a la raíz cuadrada del coeficiente de fiabilidad:

$$\rho_{xv} = \sqrt{\rho_{xx'}} = \frac{\sigma_v}{\sigma_x}$$

A la hora de interpretar el valor del coeficiente de fiabilidad no existe un criterio único y universalmente aceptado como adecuado. Evidentemente, valores cercanos a 0 denotarán una alta proporción de la varianza de los errores en la varianza de las puntuaciones empíricas, y por lo tanto, pondrán de manifiesto que el instrumento utilizado no es fiable, mientras que valores cercanos a 1 mostrarán una baja proporción de la varianza de los errores en la varianza de las puntuaciones empíricas y, en consecuencia, nos permitirán interpretar que el test utilizado es fiable. Ahora bien, el significado de esta varianza de error difiere con relación al tipo de estrategia que se ha utilizado para valorar la fiabilidad (estas estrategias se describen en los próximos apartados). Cohen y Swerlik (2009) proponen que si se ha utilizado la estrategia de test-retest, la varianza de error será debida fundamentalmente a las diferentes administraciones del test; si se ha utilizado la estrategia de formas paralelas, el error se puede atribuir a la construcción del test o a las diferentes administraciones, y si se ha valorado la fiabilidad a partir de la consistencia del test, la varianza de error puede deberse a la construcción del test.

Aparte de los casos extremos, la determinación del valor mínimo aceptable del coeficiente de fiabilidad depende de factores que pueden influir en este valor, como la longitud del test o el procedimiento empírico o la estrategia utilizada para su cálculo, tal como se ha comentado en el párrafo anterior. En cualquier caso, se han intentado establecer ciertos criterios generales que nos pueden servir de referencia. Así, en su texto clásico, Nunnally (1978) considera que el valor mínimo aceptable del coeficiente de fiabilidad estaría en 0,70, sobre todo en un contexto de investigación básica. En cambio, en un contexto aplicado, como el escolar o el clínico, es necesario que la fiabilidad sea más elevada, situándola por encima de 0,80 o 0,90. En estos ámbitos es necesario tener en cuenta que las consecuencias de la precisión de los instrumentos de medida utilizados pueden ser más decisivas para los sujetos evaluados (pensemos en los tests de diagnóstico clínico, o en los de inteligencia en población infantil, para determinar la necesidad de clases especiales por los niños). Murphy y Davidshofer (2005) afirman que en cualquier contexto de evaluación una fiabilidad por debajo de 0,6 se consideraría baja e inaceptable. Kapplan y Saccuzo (2009) van algo más allá y sugieren que coeficientes de fiabilidad que oscilan entre 0,7 y 0,8 son suficientemente buenos para la mayoría de las ocasiones en las que los tests se utilizan para fines de investigación.

Otros autores consideran que un coeficiente de fiabilidad muy cercano a 1 puede significar que los ítems que componen el test son redundantes al evaluar ciertos elementos o factores del constructo medido, y por lo tanto no aportan información relevante respecto a otros elementos o factores de este constructo, lo que tampoco se puede considerar como adecuado.

Sin querer establecer criterios estrictos y teniendo en consideración todo lo que se ha expuesto hasta aquí, podríamos concluir que, en general, es posible interpretar como una fiabilidad adecuada valores del coeficiente de fiabilidad dentro del intervalo de 0,70 a 0,95.

1.3. Tipos de errores de medida

Hasta este momento solo nos hemos referido a un tipo de error: el error de medida, pero hay que mencionar que este no es el único error descrito en el ámbito de la psicometría, sino que también podemos hacer referencia al error de estimación, al error de sustitución y al error de predicción.

Estos errores están relacionados con las puntuaciones de los sujetos individualmente consideradas. Así, el error de medida es, tal como lo definiremos a continuación, la diferencia entre la puntuación obtenida por un sujeto en el test y su puntuación verdadera en la característica medida por este test. Ahora bien, si consideramos los errores no individualmente sino en relación con un grupo o muestra de sujetos, podemos obtener los denominados errores típicos, que son las desviaciones típicas de estos errores calculadas a partir de las puntuaciones de todos los sujetos de la muestra.

Por lo tanto, podemos definir más formalmente estos diferentes tipos de errores, sus errores típicos asociados y las fórmulas que los expresan.

- **Error de medida**. Definimos el error de medida como la diferencia entre la puntuación empírica de un sujeto (X) y su puntuación verdadera (V).

$$e = X - V$$

  El **error típico de medida** es la desviación típica de los errores de medida, y lo podemos expresar como:

$$\sigma_e = \sigma_x\sqrt{1-\rho_{xx'}}$$

- **Error de estimación de la puntuación verdadera**. El error de estimación de la puntuación verdadera se define como la diferencia entre la puntuación verdadera de un sujeto y su puntuación verdadera pronosticada mediante el modelo de la regresión (V').

$$e = V - V'$$

  La desviación típica de estos errores de estimación se denomina **error típico de estimación** y se puede obtener con la siguiente expresión:

$$\sigma_{v.x} = \sigma_x\sqrt{1-\rho_{xx'}}\sqrt{\rho_{xx'}} = \sigma_e\sqrt{\rho_{xx'}}$$

- **Error de sustitución**. Se define el error de sustitución como la diferencia entre las puntuaciones de un sujeto en dos formas paralelas de un test o, dicho de otra manera, el error que se comete al sustituir la puntuación

de un sujeto en un test (X_1), por la puntuación obtenida en una forma paralela de este mismo test (X_2).

$$e = X_1 - X_2$$

Se denomina **error típico de sustitución** a la desviación típica de los errores de sustitución, y lo podemos expresar del siguiente modo:

$$\sigma_{e(s)} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{2}$$

- **Error de predicción**. El error de predicción podemos definirlo como la diferencia entre la puntuación de un sujeto en un test (X_1) y la puntuación pronosticada en este test ($X'_{1,}$) a partir de una forma paralela X_2. Sería el error que cometeríamos si sustituyéramos la puntuación de un sujeto en un test por la puntuación pronosticada a partir de una forma paralela de este test.

$$e = X_1 - X'_1$$

En este sentido, X'_1 será la puntuación pronosticada mediante la recta de regresión de X_1 sobre X_2, y la podemos expresar a partir del modelo lineal general adaptado a este contexto como:

$$X'_1 = \rho_{12} \frac{\sigma_1}{\sigma_2} (X_2 - \bar{X}_2) + \bar{X}_1$$

Definimos el **error típico de predicción** como la desviación típica de los errores de predicción, y lo podemos expresar como:

$$\sigma_{e(p)} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{1 + \rho_{xx'}}$$

2. Equivalencia de las medidas: Método de las formas paralelas

De la definición y las fórmulas del coeficiente de fiabilidad no se puede extraer directamente ningún procedimiento que nos permita calcular su valor para una determinada muestra de sujetos. Así, la estimación empírica del valor del coeficiente de fiabilidad hay que obtenerla mediante alguna estrategia que nos permita o bien comparar las puntuaciones de los mismos sujetos en dos administraciones del mismo test o en dos formas paralelas del test, o bien analizar las puntuaciones de un grupo de sujetos en los diferentes ítems del test.

De los procedimientos empíricos para la obtención del coeficiente de fiabilidad, el que se deriva directamente de la TCT es el llamado **método de las formas paralelas**, que consiste en el cálculo del coeficiente de correlación de Pearson entre las puntuaciones de una amplia muestra de sujetos representativa de la población diana del test, en dos formas paralelas de un test previamente obtenidas. Si, tal como se puede derivar de la definición de formas paralelas de un test, estas miden exactamente el mismo constructo, exactamente con la misma precisión, las diferencias que podremos observar entre las puntuaciones de unos mismos sujetos en las dos formas deben ser consecuencia de los errores de medida del test, y por lo tanto este procedimiento nos proporcionará un indicador adecuado de la magnitud de estos errores de medida, o sea, de la precisión o fiabilidad del test.

De hecho, este indicador también será representativo del grado de equivalencia de las dos formas paralelas del test, y en este sentido también recibe el nombre de **coeficiente de equivalencia**.

Su fórmula será la del coeficiente de correlación de Pearson aplicado a este caso:

$$r_{xx'} = r_{x_1x_2} = \frac{n\sum x_1x_2 - \sum x_1 \sum x_2}{\sqrt{[n\sum x_1^2 - (\sum x_1)^2][n\sum x_2^2 - (\sum x_2)^2]}}$$

Lectura de la fórmula

$r_{xx'}$: Coeficiente de fiabilidad del test.

rx_1x_2: Coeficiente de correlación de Pearson.

x_1 y x_2: Puntuaciones obtenidas por los sujetos en cada una de las dos formas paralelas del test.

Hemos designado el coeficiente de fiabilidad del test con la notación de $r_{xx'}$ y no la de $\rho_{xx'}$ porque estamos obteniendo este coeficiente de fiabilidad para una muestra concreta de sujetos, y por lo tanto nos situamos a un nivel empírico y no a un nivel teórico, como hacíamos en los aparatos anteriores al presentar las bases y definiciones de la fiabilidad y de sus errores. Esta notación, que podemos definir como muestral[1], es la que seguiremos utilizando en los

[1] Aplicada a una muestra concreta de sujetos.

siguientes apartados, dado que en todos ellos se presentan los procedimientos empíricos para la obtención del coeficiente de fiabilidad y otros indicadores relacionados con él.

La mayor dificultad para la aplicación del método de las formas paralelas recae en la elaboración de estas dos versiones de un test. A menudo resulta realmente muy difícil construir dos tests que estén formados por ítems que puedan ser emparejados en función de su total equivalencia, como requiere el concepto y la definición de formas paralelas.

Los otros procedimientos para la obtención del coeficiente de fiabilidad se derivan de las dos vertientes que podemos considerar como inherentes al concepto de fiabilidad: la estabilidad temporal y la consistencia interna.

3. Estabilidad de las medidas: Método test-retest

Un instrumento de medida fiable debe proporcionar valores estables en diferentes medidas de los mismos sujetos secuencialmente obtenidas. Así, si medimos un rasgo de personalidad supuestamente estable de unos mismos sujetos en dos diferentes ocasiones con el mismo test, el coeficiente de correlación entre sus puntuaciones será un buen indicador de esta estabilidad del test, y por lo tanto, de su fiabilidad.

Este método de obtención del coeficiente de fiabilidad de un test se denomina **método test-retest** y consiste en la aplicación del test a una misma muestra de sujetos en dos ocasiones diferentes. Se calcula a partir del valor del coeficiente de correlación de Pearson entre las puntuaciones de los sujetos en estas dos ocasiones. La fórmula es exactamente la misma aplicada en el apartado anterior para el caso de las formas paralelas, únicamente con la diferencia de que en el test-retest, x_1 y x_2 son las puntuaciones obtenidas por los sujetos en las dos administraciones del test.

La ventaja de este método es que no requiere dos formas diferentes del test (con todas las dificultades que esto implica) y su principal inconveniente se deriva en que hay que administrar dos veces el mismo test a los mismos sujetos. Este hecho puede suponer factores que distorsionen las puntuaciones de los sujetos en la segunda administración. Así, si los sujetos todavía recuerdan el contenido del test, seguramente su rendimiento se verá mejorado respecto a la primera administración. En este sentido, un factor crucial para una correcta aplicación de este método es determinar el intervalo temporal que hay que dejar entre las dos administraciones del test. Este intervalo temporal no puede ser ni demasiado corto como para provocar los efectos comentados anteriormente, ni demasiado amplio como para que se puedan dar cambios naturales (madurativos, evolutivos o circunstanciales) del rasgo o constructo medido que modifiquen las puntuaciones de los sujetos en esta segunda administración del test.

4. Consistencia interna

Tal como se ha dicho anteriormente, un instrumento de medida fiable se caracteriza por una elevada estabilidad temporal y por una adecuada consistencia interna. La consistencia interna hace referencia al grado en que cada una de las partes de las que se compone el instrumento es equivalente al resto. Este principio aplicado al caso de los tests vendrá determinado por el grado en el que cada ítem, como parte básica constitutiva de este, muestra una equivalencia adecuada con el resto de los ítems, o sea, que mide con el mismo grado el constructo medido. Así, si hay una elevada equivalencia entre los ítems del test, es de suponer que las respuestas de los sujetos a estos diferentes ítems estarán altamente correlacionadas, y las diferentes partes en las que podamos dividir el test también mostrarán esta elevada covariación.

Por poner un ejemplo, pese a las evidentes diferencias que se pueden establecer con el caso de los tests, la consistencia interna de una cinta métrica queda garantizada si cada una de sus partes (supongamos los diferentes centímetros que la componen) es equivalente al resto. Así, podremos dividir la cinta en diferentes partes iguales (dos, tres etc.), y cada una de ellas medirá exactamente la misma distancia. Evidentemente, esta exactitud en la consistencia interna de la cinta métrica es prácticamente imposible de lograr en el caso de los tests, pero el ejemplo puede servir para situar adecuadamente el concepto de consistencia interna referido al caso de la construcción de instrumentos de medida en psicología.

4.1. Método de las dos mitades

Puede derivarse fácilmente, a partir de lo que se ha expuesto en el apartado anterior, que si dividimos un test en dos mitades, estas deben ser equivalentes para garantizar una adecuada consistencia interna. El grado de equivalencia de las dos mitades se puede evaluar calculando la correlación entre las puntuaciones de los sujetos en estas dos mitades. Así, la correlación entre las puntuaciones de un grupo de sujetos en las dos mitades en las que podemos dividir un test será un indicador del grado de consistencia interna de este, y por lo tanto de su fiabilidad. Este es el principio en el que se basa el **método de las dos mitades**, que presenta la ventaja respecto a los métodos anteriores de que solo requiere una sola aplicación del test a una muestra de sujetos.

A la hora de decidir cómo realizar esta partición del test en dos mitades, hay que tener en cuenta que si lo hacemos, por ejemplo, dejando los primeros ítems en una mitad y los últimos en la otra, pueden ponerse en juego factores que alteren la equivalencia del rendimiento de los sujetos en las dos mitades. Así, es conocido que los sujetos suelen prestar más atención a los primeros ítems de un test, con la consecuente mejora de su rendimiento, o a una mayor sinceridad en sus respuestas. Estos posibles factores incidirían en una falta de consistencia interna del test, no producto de sus errores de medida aleatorios

propios de la fiabilidad del test, sino de errores de medida sistemáticos independientes de esta fiabilidad, dado que una de las dos mitades del test se vería favorecida por un mejor o más cuidadoso rendimiento de los sujetos.

Para evitar factores como los comentados anteriormente, habitualmente se divide el test en dos mitades, dejando los ítems pares en una mitad y los impares en la otra. Con este procedimiento se evitan buena parte de estos factores y se garantiza de manera más probable la equivalencia entre las dos mitades.

4.1.1. Spearman-Brown

Como se comentará con más detalle en los próximos apartados, el número de ítems que componen un test incide en su fiabilidad. Así, siendo constantes otros factores, cuantos más ítems contiene un test más elevada es su fiabilidad. Este efecto de la longitud de un test sobre el coeficiente de fiabilidad hay que tenerlo presente al aplicar el método de las dos mitades. Por lo tanto, si calculamos el coeficiente de correlación entre el total de las puntuaciones de los sujetos en los ítems pares por un lado y por otro el total de sus puntuaciones en los ítems impares, y a partir de este coeficiente de correlación cuantificáramos la fiabilidad del test, este valor estaría negativamente sesgado, dado que lo calcularíamos a partir de la correlación entre la mitad del número total de ítems del test. Este hecho supone que hay que realizar una corrección de este coeficiente de correlación para obtener el coeficiente de fiabilidad de la totalidad del test. Esta corrección se denomina de Spearman-Brown y es un caso concreto de la fórmula del mismo nombre que se aplica para obtener la fiabilidad de un test una vez este se ha alargado o acortado, añadiendo o eliminando una determinada cantidad de ítems:

La fórmula para la obtención del coeficiente de fiabilidad de un test a partir del método de las dos mitades con la corrección de Spearman-Brown es:

$$r_{xx'} = \frac{2r_{pi}}{1+r_{pi}}$$

Lectura de la fórmula

$r_{xx'}$: Coeficiente de fiabilidad del test.

r_{pi}: Coeficiente de correlación de Pearson entre el sumatorio de las puntuaciones de los ítems pares y las de los ítems impares.

Ejemplo

En la tabla siguiente tenemos las puntuaciones de ocho sujetos en un test de seis ítems dicotómicos:

Tabla 1

Sujetos	Ítems					
	1	2	3	4	5	6
A	1	1	1	1	0	1
B	0	1	1	1	1	0
C	1	1	0	1	1	0
D	1	1	1	1	1	1

Sujetos	Ítems					
E	1	1	1	1	1	1
F	0	1	1	0	0	0
G	0	1	1	0	1	0
H	1	0	1	0	0	0

Para obtener el coeficiente de fiabilidad del test por el método de las dos mitades, en primer lugar calculamos el sumatorio de las puntuaciones en los ítems pares para cada sujeto por un lado, y por otro el sumatorio de sus ítems impares, y obtenemos el coeficiente de correlación entre estas dos distribuciones de valores:

Tabla 2

Sujetos	Ítems pares	Ítems impares
A	3	2
B	2	2
C	2	2
D	3	3
E	3	3
F	1	1
G	1	2
H	0	2

Aplicando la fórmula del coeficiente de correlación de Pearson entre los ítems pares y los impares obtenemos un coeficiente igual a 0,62 ($r_{pi} = 0{,}62$).

Una vez calculado este valor, aplicamos la fórmula de Spearman Brown para obtener el coeficiente de fiabilidad del test:

$$r_{xx'} = \frac{2r_{pi}}{1+r_{pi}} = \frac{2 \times 0{,}62}{1+0{,}62} = 0{,}76$$

El coeficiente de fiabilidad del test es de 0,76.

4.1.2. Rulon

La fórmula de Rulon (1939) para calcular la fiabilidad de un test también parte de su división en dos mitades. Se basa en el supuesto de que si las dos mitades son paralelas, las puntuaciones de los sujetos en cada una de ellas solo pueden diferir como consecuencia de los errores aleatorios. Por lo tanto, la varianza de las diferencias entre estas dos mitades será una estimación de la varianza de los errores y podremos sustituir la varianza de los errores de la fórmula del coeficiente de fiabilidad derivada de la TCT por la varianza de las diferencias.

Así, la fórmula de Rulon es la siguiente:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2}$$

Lectura de la fórmula

S_d^2: Varianza de las diferencias entre las puntuaciones de los sujetos en las dos mitades del test.

S_x^2: Varianza de las puntuaciones totales de los sujetos en el test.

Ejemplo

Si aplicamos la fórmula de Rulon al ejemplo del apartado anterior, necesitaremos calcular:

Tabla 3

Sujetos	Ítems pares	Ítems impares	Diferencia P-I	Total
A	3	2	1	5
B	2	2	0	4
C	2	2	0	4
D	3	3	0	6
E	3	3	0	6
F	1	1	0	2
G	1	2	–1	3
H	0	2	–2	2

Después de calcular la varianza de las diferencias $\left(S_d^2 = 0{,}6875\right)$ y la varianza de las puntuaciones totales $\left(S_x^2 = 2{,}25\right)$, podemos aplicar la fórmula de Rulon:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2} = 1 - \frac{0{,}6875}{2{,}25} = 0{,}69$$

El coeficiente de fiabilidad del test obtenido con la fórmula de Rulon es de 0,69.

4.1.3. Gutman-Flanagan

Tanto Flanagan (1937) como Gutman (1945) obtuvieron una fórmula equivalente a la de Rulon a partir de las varianzas de los ítems pares e impares. Se basa en el mismo principio que el anterior, pero resulta más sencilla de obtener:

$$r_{xx'} = 2\left(1 - \frac{S_p^2 + S_i^2}{S_x^2}\right)$$

Lectura de la fórmula

S_p^2: Varianza de las puntuaciones de los sujetos en los ítems pares del test.

S_i^2: Varianza de las puntuaciones de los sujetos en los ítems impares del test.

S_x^2: Varianza de las puntuaciones totales de los sujetos en el test.

Ejemplo

Aplicando la fórmula de Gutman-Flanagan al ejemplo anterior, tenemos:

Tabla 4

Sujetos	Ítems pares	Ítems impares	Total
A	3	2	5
B	2	2	4
C	2	2	4

Sujetos	Ítems pares	Ítems impares	Total
D	3	3	6
E	3	3	6
F	1	1	2
G	1	2	3
H	0	2	2

Calculando las diferentes varianzas obtenemos:

Varianza de las puntuaciones en los ítems pares: $S_p^2 = 1{,}11$

Varianza de las puntuaciones en los ítems impares: $S_i^2 = 0{,}36$

Varianza de las puntuaciones totales en el test: $S_x^2 = 2{,}25$

y por lo tanto:

$$r_{xx'} = 2\left(1 - \frac{S_p^2 + S_i^2}{S_x^2}\right) = 2\left(1 - \frac{1{,}11 + 0{,}36}{2{,}25}\right) = 0{,}69$$

Como podemos observar, el valor del coeficiente de fiabilidad, calculado a partir de la expresión de Gutman-Flanagan, es exactamente igual al obtenido con la fórmula de Rulon, como no podía ser de otra manera, dado que, como hemos dicho, las dos fórmulas son equivalentes. Tanto una como la otra proporcionan un coeficiente de fiabilidad del test igual a 0,69.

4.2. Covariación entre los ítems

Si, tal como se ha comentado anteriormente, la consistencia interna de un test hace referencia al grado en el que cada una de las partes o ítems de los que se compone es equivalente al resto, la covariación entre estos ítems también nos proporcionará un adecuado indicador de esta consistencia interna. De hecho, este procedimiento es una extensión de los procedimientos anteriores de la división del test en dos mitades, al caso límite de dividirlo en tantas partes como ítems lo componen. Así, cada ítem representará una parte equivalente del conjunto de todos ellos, es decir, del test o escala total. Del mismo modo que las dos partes del test deben mantener una elevada correlación entre ellos para garantizar la misma consistencia interna del conjunto, cada ítem también ha de mostrar una covariación adecuada con el resto de los ítems.

4.2.1. Coeficiente alfa de Cronbach

El coeficiente alfa (α) de Cronbach (1951) expresa la consistencia interna de un test a partir de la covariación entre sus ítems. Cuanto más elevada sea la proporción de la covariación entre estos ítems respecto a la varianza total del test, más elevado será el valor del coeficiente alfa (α) de Cronbach, y más elevada su fiabilidad.

Existen diferentes fórmulas para obtener el valor del coeficiente α, la más ampliamente utilizada de las cuales es la que se deriva del cálculo de las varianzas de cada ítem y de la varianza de las puntuaciones totales en el test.

Esta fórmula es la siguiente:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{j=1}^{n} S_j^2}{S_x^2}\right]$$

Lectura de la fórmula

n: Número de ítems del test.

$\sum_{j=1}^{n} S_j^2$: Sumatorio de las varianzas de los *n* ítems.

S_x^2: Varianza de las puntuaciones totales en el test.

En nuestro ejemplo:

Tabla 5

Sujetos	Ítems						
	1	2	3	4	5	6	x
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2
Varianzas	0,234	0,109	0,109	0,234	0,234	0,234	2,25

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{j=1}^{n} S_j^2}{S_x^2}\right] = \frac{6}{5}\left(1 - \frac{0{,}234 + 0{,}109 + 0{,}109 + 0{,}234 + 0{,}234 + 0{,}234}{2{,}25}\right) = 0{,}583$$

Por otro lado, la fórmula del coeficiente alfa que se deriva directamente de la covarianza entre los diferentes ítems viene expresada por:

$$\alpha = \frac{n}{n-1}\left[\frac{\sum\sum_{j \neq k}^{n} \mathrm{cov}(j,k)}{S_x^2}\right]$$

Lectura de la fórmula

n: Número de ítems del test.

$\sum\sum \mathrm{cov}(j,k)$: Sumatorio de las covarianzas de los *n* ítems.

S_x^2: Varianza de las puntuaciones en el test.

Así, en el ejemplo expuesto anteriormente, calcularemos la varianza de las puntuaciones totales en el test (*x*), y las covarianzas entre los diferentes ítems:

Tabla 6

Sujetos	Ítems						
	1	2	3	4	5	6	x
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2

Las covarianzas entre los 6 ítems son:

Tabla 7

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Ítem 1		–0,047	–0,047	0,109	–0,016	0,141
Ítem 2	–0,047		–0,016	0,078	0,078	0,047
Ítem 3	–0,047	–0,016		–0,047	–0,047	0,047
Ítem 4	0,109	0,078	–0,047		0,109	0,141
Ítem 5	–0,016	0,078	–0,047	0,109		0,016
Ítem 6	0,141	0,047	0,047	0,141	0,016	

y su sumatorio:

$$\sum\sum \mathrm{cov}(j,k) = -0{,}047 + -0{,}047 + 0{,}109 + \ldots\ldots + 0{,}047 + 0{,}141 + 0{,}016 = 1{,}094$$

mientras que la varianza de las puntuaciones totales: $S_x^2 = 2{,}25$

Por lo tanto:

$$\alpha = \frac{n}{n-1}\left[\frac{\sum\sum_{j\neq k}^{n}\mathrm{cov}(j,k)}{S_x^2}\right] = \frac{6}{5}\left(\frac{1{,}094}{2{,}25}\right) = 0{,}583$$

La fórmula del coeficiente α también se puede expresar en función del cociente entre la media de las covarianzas y la media de las varianzas de los diferentes ítems del test. Este cociente, que designamos como r_1, constituye una estimación de la fiabilidad de cada ítem. En este sentido, la fórmula del coeficiente α a partir de r_1 es una aplicación de la corrección de Spearman-Brown, que

hemos comentado para el caso de las dos mitades, a partir de la estimación de la fiabilidad de cada ítem, teniendo en cuenta que si tenemos *n* ítems es como si hubiéramos alargado *n* veces el ítem inicial.

$$\alpha = \frac{n(r_1)}{1+(n-1)r_1}$$

Para nuestro ejemplo, la media de las covarianzas es 1,094/30, mientras que la media de las varianzas es 1,154/6.

Por lo tanto, $r_1 = \frac{1{,}094/30}{1{,}154/6} = 0{,}189$ y el valor de α:

$$\alpha = \frac{n(r_1)}{1+(n-1)r_1} = \frac{6 \times 0{,}189}{1+(5 \times 0{,}189)} = 0{,}583$$

Como podemos observar, y como no podía ser de otra manera, todas las diferentes fórmulas de cálculo del coeficiente α de Cronbach aplicadas a los datos de nuestro ejemplo nos proporcionan el mismo valor.

4.2.2. Inferencias sobre α

Una vez hemos obtenido el valor del coeficiente alfa de Cronbach para una muestra determinada de sujetos, podemos estar interesados en comprobar su significación estadística, o en determinar entre qué valores puede fluctuar este coeficiente en la población. Por otro lado, también puede interesarnos comparar dos coeficientes alfas obtenidos en dos muestras independientes, o en la propia muestra, para determinar si la diferencia entre ellos es estadísticamente significativa.

Contraste para un solo coeficiente

Kristof (1963) y Feldt (1965) propusieron un estadístico de contraste para comprobar si un determinado valor del coeficiente alfa puede ser compatible con un cierto valor poblacional. Así, podemos analizar si este valor de alfa es estadísticamente significativo, esto es, si podemos descartar la hipótesis de que su valor en la población es cero, o si este valor difiere significativamente o no de un determinado valor previamente fijado en la población.

El estadístico de contraste es:

$$F = \frac{1-\alpha}{1-\hat{\alpha}}$$

Y se distribuye según una distribución *F* de Snedecor con $(N-1)$ y $(n-1)(N\text{-}1)$ grados de libertad.

Lectura de la fórmula

N: Número de sujetos.

n: Número de ítems.

α: Valor de alfa en la población.

$\hat{\alpha}$: Valor de alfa calculado en la muestra.

Podemos aplicar este estadístico de contraste a nuestro ejemplo del test de seis ítems que nos ha dado un valor de alfa de 0,583, obtenido en una muestra de ocho sujetos.

La hipótesis nula que plantearemos es que este coeficiente alfa es igual a cero en la población, caso más habitual y que supone la no significación estadística de este coeficiente, mientras que la alternativa supondrá la desigualdad respecto al valor cero, y por lo tanto su significación estadística.

Los pasos que se deberán seguir en este contraste serán:

Hipótesis nula: $\alpha = 0$

Hipótesis alternativa: $\alpha \neq 0$

Cálculo del estadístico de contraste:

$$F = \frac{1-\alpha}{1-\hat{\alpha}} = \frac{1-0}{1-0,583} = 2,398$$

Los valores críticos de la distribución *F* de Snedecor con 7 $(N\text{-}1)$ y 35 $((n\text{-}1)(N\text{-}1))$ grados de libertad, para un nivel de confianza del 95% y contraste bilateral, son:

$$F_{0,975(7,35)} \approx 2,62 \text{ y } F_{0,025(7,35)} \approx 0,23^{2}$$

Como el valor del estadístico de contraste obtenido (2,398) se encuentra dentro del intervalo comprendido entre los valores críticos 2,62 y 0,23, aceptamos la hipótesis nula y podemos concluir que, a partir de nuestros datos y con un nivel de confianza del 95%, no tenemos evidencia suficiente para descartar que el valor del coeficiente alfa en la población es cero, por lo que este coeficiente no es estadísticamente significativo.

Como derivación sencilla de lo que se ha expuesto en este apartado, podemos también determinar el intervalo de confianza para el valor del coeficiente alfa obtenido. En este sentido, solo hay que sustituir, en la fórmula del estadístico de contraste, los valores críticos de la distribución *F* y aislar los valores de α:

$$\frac{1-\alpha}{1-0,583} \leq 2,62$$

$$\alpha \geq 1-2,62(1-0,583) = -0,09$$

$$\frac{1-\alpha}{1-0,583} \geq 0,23$$

$$\alpha \leq 1-0,23(1-0,583) = 0,90$$

[(2)]**Nota**

$$F_{0,025(7,35)} = \frac{1}{F_{0,975(35,7)}} \approx \frac{1}{4,31} = 0,23$$

La interpretación de estos valores irá en el sentido de considerar que, con un nivel de confianza del 95%, los valores del coeficiente alfa en la población estarán comprendidos entre –0,09 y 0,90. Una vez establecido este intervalo confidencial, podríamos resolver la aceptación o no de cualquier valor del coeficiente en la hipótesis nula del contraste correspondiente. Así, si el valor del coeficiente alfa poblacional planteado en la hipótesis nula cae dentro del

intervalo de confianza, no podemos descartar la certeza de esta hipótesis nula, mientras que si no cae, podremos descartarla y aceptar la hipótesis alternativa para el nivel de confianza establecido.

En nuestro ejemplo, como el valor de cero del coeficiente está comprendido en el intervalo (–0,09 - 0,90) no podemos rechazar la hipótesis nula, tal como ya hemos visto anteriormente. Se deriva directamente de lo anterior el hecho de que si un coeficiente alfa empíricamente obtenido no se encuentra comprendido en el intervalo de confianza construido, queda determinada su significación estadística sin necesidad de realizar el contraste para una hipótesis nula igual a cero.

Contraste para dos coeficientes en muestras independientes

También podemos estar interesados en comprobar si dos coeficientes alfa obtenidos en muestras diferentes de sujetos son iguales o no. Para responder a esta cuestión aplicaremos un contraste para dos coeficientes en muestras independientes. Feldt (1969) propuso el estadístico *w*, que permite determinar si la diferencia entre los dos coeficientes es estadísticamente significativa:

$$w = \frac{1 - \hat{\alpha}_1}{1 - \hat{\alpha}_2}$$

donde *w* se distribuye según una F de Snedecor con $(N_1 - 1)$ y $(N_2 - 1)$ grados de libertad.

Lectura de la fórmula

$\hat{\alpha}_1$ y $\hat{\alpha}_2$: Coeficientes obtenidos en cada una de las dos muestras.

N_1 y N_2: Dimensiones de estas muestras.

Podemos realizar este contraste para nuestro ejemplo con un nivel de confianza del 95%, suponiendo que se ha administrado el mismo test a una muestra de 10 sujetos y que hemos obtenido un coeficiente alfa de 0,65. Realizaremos el contraste siguiendo los siguiendo pasos:

Hipótesis nula: $\hat{\alpha}_1 = \hat{\alpha}_2$

Hipótesis alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

Cálculo del estadístico de contraste: $w = \frac{1 - \hat{\alpha}_1}{1 - \hat{\alpha}_2} = \frac{1 - 0{,}583}{1 - 0{,}65} = 1{,}19$

Los valores críticos de la distribución *F* de Snedecor con 7 (N_1) y 9 (N_2) grados de libertad, para un nivel de confianza del 95% y contraste bilateral, son:

$$F_{0{,}975(7,9)} = 4{,}20 \text{ y } F_{0{,}025(7,9)} = 0{,}21$$

Como el valor del estadístico de contraste obtenido (1,19) cae dentro del intervalo comprendido entre los valores críticos (0,21 - 4,20), no tenemos suficientes evidencias para rechazar la hipótesis nula, y por lo tanto debemos concluir que la diferencia entre los dos coeficientes no es estadísticamente significativa.

Lecturas recomendadas

Woodruff y Feldt (1986) hicieron extensivo el contraste anterior al caso de más de dos coeficientes comparados simultáneamente. No nos extenderemos en esta cuestión, dado que su aplicación no es tan frecuente, pero remitimos al estudiante interesado a los textos de Muñiz (2003) y Barbero, Vila y Suárez (2003), en los que se puede encontrar su desarrollo claramente explicado.

Contraste para dos coeficientes en muestras dependientes

Es habitual que los dos coeficientes alfa obtenidos se hayan calculado a partir de la misma muestra de sujetos. Menos frecuente es que se hayan obtenido a partir de dos muestras de sujetos relacionados entre ellos por algún criterio de emparejamiento (por ejemplo, parejas de gemelos, padre-madre). No obstante, tanto en uno como en el otro supuesto denominamos el diseño como muestras dependientes. Sería el caso, por ejemplo, de aplicar un diseño experimental de medidas repetidas y administrar el mismo test a un solo grupo de sujetos en dos ocasiones diferentes. En este sentido, podríamos comparar los dos coeficientes alfa obtenidos para determinar la posible diferencia estadísticamente significativa entre ellos.

Feldt (1980) propuso un estadístico de contraste para comparar dos coeficientes α obtenidos en muestras dependientes:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}}$$

Lectura de la fórmula

t: Se distribuye según una distribución *t* de Student con *N* – 2 grados de libertad.

$\hat{\alpha}_1$ y $\hat{\alpha}_2$: Valores de los dos coeficientes alfa.

N: Número de sujetos de la muestra.

r_{12}: Correlación entre las puntuaciones de los sujetos en las dos administraciones del test.

Para ilustrar la aplicación de este contraste podemos considerar que el test utilizado en los apartados anteriores lo hemos administrado de nuevo en una segunda ocasión a la misma muestra de ocho sujetos.

En la tabla siguiente tenemos las puntuaciones de estos sujetos en las dos ocasiones en las que se les ha administrado el test.

Tabla 8

	Ocasión 1							Ocasión 2						
	Ítems							Ítems						
Sujetos	1	2	3	4	5	6	x_1	1	2	3	4	5	6	x_2
A	1	1	1	1	0	1	5	1	1	0	1	0	1	4
B	0	1	1	1	1	0	4	0	1	0	1	1	0	3
C	1	1	0	1	1	0	4	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6	1	1	0	1	1	1	5
E	1	1	1	1	1	1	6	1	1	0	1	1	1	5
F	0	1	1	0	0	0	2	0	1	0	0	0	0	1
G	0	1	1	0	1	0	3	0	1	0	0	1	0	2
H	1	0	1	0	0	0	2	1	0	0	0	0	0	1

A partir de estos datos, calculamos el coeficiente alfa para la segunda administración del texto ($\hat{\alpha}_2 = 0{,}718$), y la correlación entre las puntuaciones totales de los sujetos en estas dos ocasiones ($r_{12} = 0{,}98$).

El contraste para determinar la posible diferencia estadísticamente significativa entre los dos coeficientes alfa, con un nivel de confianza del 95%, seguirá los pasos siguientes:

Hipótesis nula: $\hat{\alpha}_1 = \hat{\alpha}_2$

Hipótesis alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

Cálculo del estadístico de contraste:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}} = \frac{(0{,}583 - 0{,}718)\sqrt{8-2}}{\sqrt{4(1-0{,}583)(1-0{,}718)(1-0{,}98)}} = -3{,}41$$

Los valores críticos de la distribución *t* de Student con 6 (*N* – 2) grados de libertad, para un nivel de confianza del 95% y contraste bilateral, son:

$$t_{0{,}975(6)} = 2{,}447 \text{ y } t_{0{,}025(6)} = -2{,}447$$

Como el estadístico de contraste obtenido (–3,41) queda fuera del intervalo entre los valores críticos (–2,447 - 2,447), podemos rechazar la hipótesis nula y aceptar la alternativa, y concluir que con un nivel de confianza del 95% la diferencia entre los dos coeficientes alfa es estadísticamente significativa.

Como en el caso de dos muestras independientes, Woodruff y Feldt (1986) también hicieron extensivo el contraste anterior, además de dos coeficientes comparados simultáneamente en muestras dependientes. Se puede consultar su desarrollo en los mismos textos citados para muestras independientes.

4.2.3. Kuder-Richardson

Algunos años antes de que Cronbach propusiera el coeficiente alfa como indicador de la consistencia interna de un test, Kuder y Richardson (1937) presentaron dos fórmulas de cálculo de este indicador, que de hecho son casos particulares de α cuando los ítems son dicotómicos. Estas dos fórmulas son conocidas como KR_{20} y KR_{21}.

Cuando los ítems de un test son dicotómicos y se codifican las dos alternativas de respuesta posibles como 0 y 1, la varianza de un ítem es igual a la proporción de ceros para la proporción de unos. Si el test es de rendimiento y las respuestas a los diferentes ítems son correctas o incorrectas, habitualmente se codifica con un 1 las respuestas correctas y con un 0 las incorrectas. En este caso, la varianza del ítem será igual a la proporción de sujetos que aciertan el ítem (p_j) por la proporción de sujetos que no lo aciertan (q_j). Igualmente, si el test es de personalidad y no hay respuestas correctas ni incorrectas, pero se codifica con un 1 los sujetos que responden "SÍ" y con un 0 los que responden "NO", la varianza del ítem será la proporción de sujetos que responden "SÍ" (p_j) por la proporción de sujetos que responden "NO" (q_j). En los dos casos $S_j^2 = p_j q_j$.

Teniendo en cuenta esta igualdad, la fórmula del KR_{20} simplemente sustituye en la del coeficiente α de Cronbach el sumatorio de las varianzas de los ítems por el sumatorio de los productos p_j por q_j:

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{S_x^2}\right)$$

En el supuesto de que todos los ítems tuvieran la misma dificultad, o de que el número de sujetos que responden "SÍ" se mantuviera constante para todos los ítems, el producto p_j por q_j sería igual para todos ellos, y su sumatorio sería igual a la media del test menos esta media al cuadrado dividida por el número de ítems (n). Esta nueva igualdad permite reformular el KR_{20} para el caso de que todos los ítems tengan la misma dificultad o el número de sujetos que responden "SÍ" se mantenga constante para todos los ítems. Esta reformulación es el KR_{21}.

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{\overline{X} - \overline{X^2}/n}{S_x^2}\right)$$

Si aplicamos estas fórmulas a nuestro ejemplo, obtenemos los resultados siguientes:

Tabla 9

Sujetos	**Ítems**						
	1	2	3	4	5	6	*x*
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2
$p_j q_j$	0,234375	0,109375	0,109375	0,234375	0,234375	0,234375	

$$\sum p_j q_j = 0{,}2343375 + 0{,}109375 + 0{,}109375 + 0{,}234375 + 0{,}234375 + 0{,}234375 = 1{,}15625$$

$$S_x^2 = 2{,}25$$

$$\overline{X} = 4$$

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{S_x^2}\right) = \frac{6}{5}\left(1 - \frac{1,15625}{2,25}\right) = 0,583$$

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{\overline{X} - \bar{X}^2 / n}{S_x^2}\right) = \frac{6}{5}\left(1 - \frac{4 - 4^2 / 6}{2,25}\right) = 0,489$$

Como podemos comprobar, el KR_{20} proporciona un resultado idéntico al del coeficiente alfa, mientras que el KR_{21} da un resultado inferior, dado que en nuestro ejemplo no todos los ítems tienen la misma dificultad. En este sentido, el cálculo del KR_{21} no sería adecuado para este caso.

5. Factores que afectan a la fiabilidad

La fiabilidad de un test depende de factores como la variabilidad de las puntuaciones del test, el número total de ítems del test o las características de los ítems que lo componen. A continuación, se tratarán estos tres aspectos: variabilidad, longitud y características de los ítems.

a) Variabilidad

En los apartados previos se ha abordado la fiabilidad a partir del cálculo del coeficiente de correlación entre dos tests paralelos, entre dos administraciones del test en dos momentos temporales diferentes o entre diferentes partes del test. Sin embargo, hay que tener en cuenta que el coeficiente de correlación es sensible al rango y variabilidad de los datos. Lo que se observa es que cuando mantenemos el resto de los factores constantes, si se aumenta la variabilidad de los datos, el coeficiente de correlación aumenta. Por esta razón, en aquellos casos en los que exista una alta variabilidad en las puntuaciones del test, el coeficiente de fiabilidad será mayor. De esto se desprende que un test no tiene un coeficiente de fiabilidad único y fijo, sino que depende de las características de la muestra sobre la que se calcula. Así, por el contrario, si la muestra es homogénea y las puntuaciones empíricas que se obtienen presentan una baja variabilidad, el coeficiente de fiabilidad será menor.

En este punto vale la pena recordar las palabras de Crocker y Algina (1986) cuando dicen que no se puede afirmar que un test es fiable o no, sino que la fiabilidad es una propiedad de las puntuaciones obtenidas en el test a partir de una muestra particular de individuos.

b) Longitud

Otro de los factores que afectan a la fiabilidad es la longitud del test. Así, la fiabilidad depende del número de ítems que presente el test. La lógica de esta afirmación subyace en que cuantos más ítems se utilicen para medir un constructo, mejor podrá ser valorado este y menor será el error de medida que se cometerá al valorar la puntuación verdadera del sujeto. Por ello, siempre que se aumente el número de ítems de un test (siempre que estos sean ítems representativos del constructo), la fiabilidad aumentará. Para saber la fiabilidad de un test en caso de que aumente o disminuya su número de ítems, se utiliza la fórmula de Spearman Brown, también conocida como profecía de Spearman Brown.

$$R_{xx} = \frac{k\, r_{xx}}{1+(k-1)r_{xx}}$$

Lectura de la fórmula

R_{xx}: Nuevo coeficiente de fiabilidad del test alargado o acortado.

r_{xx}: Coeficiente de fiabilidad del test original.

k: Número de veces que se alarga o se acorta el test.

De este modo, k vendrá dado por el cociente entre el número de ítems finales (n_f) del test dividido por el número de ítems iniciales (n_i) del test:

$$k = \frac{n_f}{n_i}$$

Si se añaden ítems a un test, k siempre será superior a 1, mientras que si se acorta el test (eliminamos ítems a los ya existentes) k será inferior a 1.

Ejemplo

Si un test de 25 ítems presenta una fiabilidad de 0,65 y le añadimos 10 ítems paralelos, ¿cuál será su fiabilidad?

$$k = \frac{35}{25} = 1{,}4$$

k indica que hay que alargar 1,4 veces la longitud del test. Si se sustituye este valor en la fórmula:

$$R_{xx} = \frac{1{,}4 \cdot 0{,}65}{1+(1{,}4-1)\cdot 0{,}65} = 0{,}72$$

Se obtiene que el nuevo coeficiente de fiabilidad del test alargado será de 0,72.

También la pregunta anterior se puede invertir y plantearse cuántos ítems debería tener el test para lograr una determinada fiabilidad. En este caso, habría que aislar k de la fórmula:

$$k = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})}$$

Efectivamente, si ahora la pregunta fuera cuántos ítems hay que añadir para conseguir una fiabilidad de 0,72 a un test de 25 ítems que presenta una fiabilidad de 0,65, se aplicaría la fórmula anterior para conocer cuántas veces habría que aumentar el test:

$$k = \frac{0{,}72 \cdot (1-0{,}65)}{0{,}65 \cdot (1-0{,}72)} = 1{,}4$$

Lo que permitirá saber el número de ítems que hay que añadir:

$$k \cdot n_i - n_i$$

$$1{,}4 \cdot 25 - 25 = 10$$

El test final tendría 35 ítems (1,4 × 25 = 35), por lo que habría que añadir 10 ítems a los 25 iniciales para conseguir una fiabilidad de 0,72.

Hay que tener presente dos aspectos: el primero es que a pesar de que la fiabilidad de un test aumentará siempre que aumentemos el número de ítems, este aumento no es directamente proporcional. Para ver el efecto del aumento que se produce en la fiabilidad en diferentes valores de k, hay que fijarse en la

figura 1, en la que en las abscisas se representan los diferentes valores de *k* (el número de veces que se ha alargado el test) y en las ordenadas el coeficiente de fiabilidad que se obtendría.

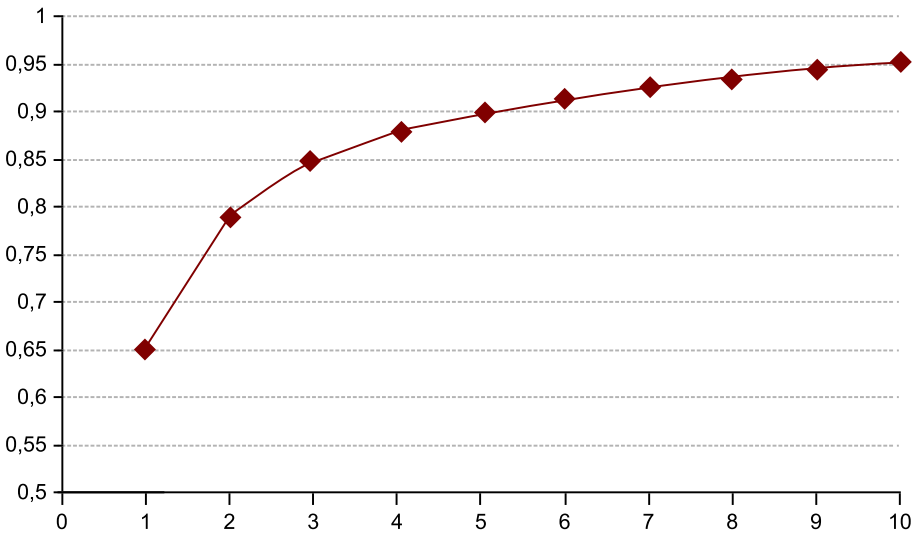

Figura 1. Relación entre el coeficiente de fiabilidad y el aumento de ítems en un test

El segundo aspecto que hay que tener en cuenta es que, aunque siempre podemos conseguir un aumento de la fiabilidad aumentando el número de ítems del test, se deben valorar los aspectos de fatiga que supone responder a un instrumento con muchos ítems. Por ello, se puede pensar en aumentar el número de ítems si el test original tiene relativamente pocos ítems y una fiabilidad que no llega a ser del todo adecuada. Pero si el test ya presenta un número considerable de ítems o su fiabilidad dista mucho ser adecuada, quizá habría que seleccionar otro test para medir el constructo o construir un nuevo test.

c) Características de los ítems

Cada ítem del test contribuye de manera específica a la fiabilidad o consistencia interna del test. Una manera de comprobarlo es calcular el coeficiente alfa de Cronbach eliminando del cálculo la puntuación del ítem. Si el ítem contribuye de manera positiva a la consistencia interna del test, al eliminarlo del test, el valor del coeficiente alfa de Cronbach se verá alterado a la baja (si eliminamos el ítem, el test pierde consistencia interna). Al contrario, si observamos que al eliminar el ítem el coeficiente alfa de Cronbach aumenta, esto indicará que el ítem no contribuye de manera positiva a la consistencia interna.

> Por ejemplo, si un test con seis ítems presenta una consistencia interna de 0,76 y se calcula de nuevo la consistencia interna eliminando la puntuación del ítem en cada caso, habría que valorar que en todos los casos la consistencia interna fuera inferior a la del conjunto del test.

En la tabla siguiente se muestra un ejemplo:

Tabla 10

Ítems	Alfa de Cronbach sin el ítem
Ítem 1	0,72
Ítem 2	0,74
Ítem 3	0,80
Ítem 4	0,71
Ítem 5	0,70
Ítem 6	0,75

En la tabla anterior, se observa que cuando se elimina un ítem el valor del coeficiente alfa de Cronbach disminuye (ítems 1, 2, 4, 5 y 6), lo que indica que el ítem contribuye de manera favorable a la consistencia interna del test. No obstante, cuando se elimina el ítem 3 el valor del coeficiente alfa de Cronbach aumenta, lo que indica que el test presentaría una mejor consistencia interna sin este ítem.

6. Estimación de la puntuación verdadera

Recordemos que según la teoría clásica la puntuación empírica del sujeto es igual a la puntuación verdadera más el error aleatorio de medida. Conocer la precisión con la que se mide el constructo permite calcular la cantidad de error que está afectando a la puntuación empírica. Así, el hecho de conocer la fiabilidad del instrumento permite estimar la puntuación verdadera del sujeto. No obstante, debe tenerse en cuenta que no se puede calcular exactamente la puntuación verdadera del sujeto, aunque sí estimarla a partir del cálculo de un intervalo de confianza.

Fundamentalmente se utilizan dos procedimientos para valorar esta puntuación verdadera:

- La estimación que asume la distribución normal del error aleatorio.
- La estimación a partir del modelo de regresión lineal.

6.1. Estimación de la puntuación verdadera a partir de la distribución normal del error aleatorio

Este procedimiento asume que el error se distribuye según la ley normal con media 0 y varianza S_e^2. A partir del cálculo de un intervalo de confianza se obtienen los límites inferior y superior entre los que se encontrará la puntuación verdadera (V) del sujeto con un determinado nivel de confianza (o bien asumiendo un determinado riesgo alfa). Los pasos que habría que seguir son los siguientes:

1. Calcular el error típico de medida (S_e) que viene dado por la expresión siguiente:

$$S_e = S_x\sqrt{1 - r_{xx}}$$

Lectura de la fórmula

S_x: Desviación típica de las puntuaciones del test.

r_{xx}: Coeficiente de fiabilidad obtenido.

2. Buscar el valor $Z_{\alpha/2}$ para el nivel de confianza con el que se quiera trabajar. Si se ha fijado el nivel de confianza al 95% (o bien el riesgo de error al 5%), el valor $Z_{\alpha/2}$ que corresponde según las tablas de la distribución normal es de 1,96.

3. Calcular el error máximo de medida ($E_{máx}$) que se está dispuesto a asumir:

$$E_{máx} = Z_{\alpha/2} \cdot S_e$$

4. Calcular el intervalo de confianza en el que se encontrará la puntuación verdadera del sujeto a partir de la expresión siguiente:

$$IC = X \pm E_{máx}$$

Ejemplo

Se ha administrado un test a una muestra de 300 sujetos. La puntuación total del test presenta una media de 15,6 puntos, una desviación típica de 5,4 y un coeficiente de fiabilidad de 0,77. Si se quiere trabajar con un nivel de confianza del 95%, ¿entre qué valores se situaría la puntuación verdadera de un sujeto que ha obtenido una puntuación empírica de 18?

Error típico de medida: $S_e = 5{,}4\sqrt{1-0{,}77} = 2{,}59$

Un nivel de confianza del 95% corresponde a una $z_{\alpha/2} = 1{,}96$

Error máximo:

$$E_{máx} = 1{,}96 \cdot 2{,}59 = 5{,}08$$

$$IC = 18 \pm 5{,}08$$

$$12{,}92 \leq V \leq 23{,}08$$

La puntuación verdadera del sujeto oscilará entre los valores 12,92 y 23,08, con un nivel de confianza del 95%.

Hay que tener en cuenta que la precisión de la medida, es decir, la fiabilidad del instrumento que utilizamos para medir el constructo, tiene un efecto directo en la amplitud del intervalo construido. Cuando utilizamos un instrumento con una fiabilidad alta, la estimación de la puntuación verdadera del sujeto será más precisa, lo que implica que la amplitud del intervalo será menor. Sucede lo contrario cuando se utilizan instrumentos menos fiables: la precisión con la cual podemos calcular la puntuación verdadera del sujeto es menor y el intervalo construido es más amplio.

6.2. Estimación de la puntuación verdadera a partir del modelo de regresión lineal

La puntuación empírica es una puntuación sesgada (teóricamente se observa una correlación positiva entre la puntuación empírica y el error de medida), por lo que la estimación que se pueda hacer a partir de esta también presentará sesgo. Por este motivo, otra posibilidad es primero valorar la puntuación verdadera y después calcular el intervalo de confianza a partir de la puntuación verdadera obtenida.

Si nos basamos en el supuesto de que la puntuación verdadera es igual a la media de la puntuación empírica, podemos hacer una estimación de la puntuación verdadera a partir de la ecuación de regresión siguiente:

$$V' = r_{xx}\,(X - \overline{X}) + \overline{X}$$

Lectura de la fórmula

V': Puntuación verdadera pronosticada.

r_{xx}: Coeficiente de fiabilidad del test.

X: Puntuación empírica obtenida por el sujeto.

$\overline{X}$: Media de las puntuaciones del test.

Una vez estimada la puntuación verdadera, se pasaría a calcular el intervalo de confianza para establecer con un determinado nivel de confianza entre qué valores se situaría la puntuación verdadera del sujeto. En este caso se utilizará el error típico de estimación (desviación típica de la diferencia entre la puntuación verdadera y la puntuación verdadera pronosticada ($V - V'$)):

$$S_{vx} = S_x\sqrt{1 - r_{xx}}\sqrt{r_{xx}} = S_e\sqrt{r_{xx}}$$

Lectura de la fórmula

S_x: Desviación típica de las puntuaciones del test.

r_{xx}: Coeficiente de fiabilidad del test.

S_e: Error típico de medida.

Continuando con este ejemplo, a continuación se muestra cómo calcular primero la puntuación verdadera pronosticada y después el intervalo de confianza que indicará entre qué valores se situará la puntuación verdadera del sujeto con un nivel de confianza del 95%.

En primer lugar se calcula la puntuación verdadera:

$$V' = 0{,}77 \cdot (18 - 15{,}6) + 15{,}6 = 17{,}45$$

Se obtiene el error típico de estimación a partir del error típico de medida calculado en el ejemplo anterior y el coeficiente de fiabilidad:

$$S_{vx} = 2{,}59\sqrt{0{,}77} = 2{,}27$$

A continuación se calcula el error máximo:

$$E_{máx} = 1{,}96 \cdot 2{,}27 = 4{,}45$$

El intervalo de confianza quedará construido a partir de la estimación de la puntuación verdadera y el error máximo obtenido.

$$IC = 17{,}45 \pm 4{,}45$$

$$13{,}00 \leq V \leq 21{,}90$$

El resultado del intervalo, ya sea a partir de la estimación basada a partir de la distribución normal de los errores o a partir del modelo de regresión lineal, se interpreta del mismo modo. En los dos casos se puede afirmar que existe una probabilidad del 95% de que un sujeto que haya tenido una puntuación empírica de 18 puntos en el test sitúe su puntuación verdadera entre 13 y 23 (o entre 13 y 22, si la estimación se ha realizado a partir del modelo de regresión lineal).

7. Fiabilidad de los tests referidos al criterio

7.1. Conceptos básicos

Los tests a los que se ha hecho referencia hasta el momento son los llamados tests referidos a la norma (TRN). Tal como se ha expuesto en el módulo "Aproximación histórica y conceptos básicos de la psicometría", estos tests tienen como objetivo escalar los sujetos en función de las puntuaciones obtenidas en un test que mide un determinado rasgo o variable psicológica. En este apartado nos centraremos en otro tipo de tests, los tests referidos al criterio (TRC).

Estos tipos de tests empiezan a hacerse populares hacia la década de los años setenta del siglo XX, cuando consiguen hacerse un lugar dentro de la teoría de los tests para evaluar los estándares de rendimiento. Su surgimiento se explica, por un lado, por el contexto contrario que se había desarrollado, fundamentalmente en Estados Unidos sobre el uso de los tests, y por otro, por la necesidad de valorar la eficacia de los programas educativos que durante los últimos años habían surgido. En este último punto habría que evaluar el nivel de habilidad y conocimiento que los individuos alcanzaron dentro del programa educativo para demostrar su eficacia. En la actualidad, su uso está muy extendido en los campos educativo y laboral, donde lo que se persigue es evaluar la competencia que presentan los individuos en un determinado conjunto de conocimientos o criterio.

Un TRC pretende evaluar en términos absolutos el estatus que exhibe un sujeto respecto a un criterio, entendiendo como criterio un dominio de conductas, contenidos o conjunto de procesos muy definido. Para lograr este punto, primero es necesario definir el conjunto de tareas que el individuo debe ser capaz de realizar para demostrar su competencia sobre el criterio. El test se construye a partir de una serie de ítems que representan el dominio que quiere ser evaluado. La estimación del criterio se realiza a partir de la puntuación que el individuo obtiene en el test. La interpretación de esta puntuación es bastante intuitiva, dado que la proporción de ítems correctamente acertados indicará la competencia que el sujeto tiene del criterio. Por ejemplo, si se pretende valorar la capacidad de matemáticas de los alumnos de tercero de primaria, habrá que definir el dominio de conductas (de contenidos o de procesos) que los alumnos deberían poder realizar. Una vez decididos los ítems (preguntas, problemas, ejercicios, etc.) que formarán parte del test, y que representan el criterio que se pretende valorar, estos se administran a los sujetos. La puntuación que se obtenga será la estimación del criterio.

En este punto vale la pena detenerse a valorar las diferencias entre los TRN y los TRC, dado que, como se mostrará en los apartados siguientes, en la práctica la distinción no está siempre clara. Mientras que los TRN tienen como finalidad saber si la puntuación del sujeto A es superior a la del sujeto B cuando se valora X, los TRC se plantean si los sujetos A y B son capaces de resolver X.

En la tabla siguiente se resumen las características básicas de estos dos tipos de test respecto al objeto que se evalúa, el análisis de la fiabilidad que realizan, el objetivo de la evaluación y el ámbito de aplicación.

Tabla 11. Características básicas de los tests referidos a la norma y de los tests referidos al criterio

	TRN	**TRC**
Objeto que se evalúa	Mide una variable psicológica o rasgo	Mide un conjunto de conocimientos o competencias (criterio)
Análisis de la fiabilidad	A partir de las diferencias entre los sujetos	A partir del dominio que presenta el individuo sobre el tema
Objetivo de evaluación	Posición relativa de un individuo respecto al resto del grupo	Grado de conocimiento que presentan los individuos en el dominio
Ámbito de aplicación	Personalidad, actitudes, etc.	Educación, ámbito laboral, evaluación de programas, etc.

En esencia, el concepto de fiabilidad en el caso de los TRC y los TRN es idéntico: en los dos casos se pretende valorar el grado de error que se comete a la hora de hacer una medición. No obstante, los procedimientos que se siguen para valorar la fiabilidad en la teoría clásica para los TRN no son, en general, apropiados para estimar la fiabilidad de los TRC. La razón es que mientras que los TRN basan la medida de fiabilidad en la variabilidad de las puntuaciones del test (recordemos que el coeficiente de fiabilidad se ha definido como la proporción de varianza de las puntuaciones verdaderas que hay en la varianza de las puntuaciones empíricas), en los TRC esta variabilidad deja de tener importancia. La finalidad que persiguen los TRC es evaluar el nivel de conocimiento que los individuos poseen sobre un criterio, y mayoritariamente basan la medida de fiabilidad en clasificar a los sujetos de manera consistente en dos grandes categorías: aquellos que dominan el criterio y aquellos que no lo dominan. Si la clasificación que se hace de los individuos a partir de una o más administraciones del test es consistente, se puede hablar de consistencia o precisión en el proceso de medida y por lo tanto, de fiabilidad. No obstante, a diferencia de los TRN, los TRC se centran en la idea de que la puntuación del test permite hacer una interpretación en términos absolutos de la capacidad del sujeto sin tener que comparar los resultados con un grupo de referencia. Las diferencias entre los individuos (la variabilidad de la medida) dejan de te-

ner importancia para evaluar la fiabilidad del test (las diferencias en la puntuación del test entre unos y otros pueden ser pequeñas, sin que esto afecte a la fiabilidad).

Como se ha visto antes, uno de los factores que afecta a los valores de la fiabilidad es la variabilidad de las puntuaciones del test; por eso los procedimientos que hasta ahora se han presentado para valorar la fiabilidad no son adecuados para ser aplicados a los TRC. Siguiendo a Shrock y Coscarelli (2007), hay tres maneras de abordar la fiabilidad en este tipo de test: aquellos procedimientos que requieren dos aplicaciones del test para valorar la consistencia de la clasificación, aquellos que solo requieren una única aplicación y aquellos en los que entra en juego el papel de los evaluadores. En este último caso, son los propios evaluadores los que juzgan la competencia de los sujetos. En los apartados siguientes se tratarán estos diferentes procedimientos para abordar la fiabilidad en los TRC.

7.2. Índices de acuerdo que requieren dos aplicaciones del test

Se considera que un test es fiable cuando se administra el mismo test en dos ocasiones a la misma muestra o cuando se aplican dos formas paralelas del test y los resultados de las dos administraciones permiten clasificar a los sujetos dentro de la misma categoría. A continuación se presentan dos de los índices de acuerdo más utilizados: el coeficiente de Hambleton y Novick ($p_{H\text{-}N}$) y el coeficiente kappa (k).

7.2.1. Coeficiente de Hambleton y Novick

El coeficiente de Hambleton y Novick (1973) valora la fiabilidad o consistencia de las clasificaciones a partir de dos administraciones del test o de dos formas paralelas del test. Para proceder con su cálculo se tiene en cuenta, por un lado, la proporción de sujetos que son clasificados de manera consistente en una misma categoría (p_c), sea esta la de competentes o la de no competentes (aptas o no aptas, etc.), y por otro lado, la proporción de clasificaciones consistentes que se esperaría por azar (p_a) .La diferencia entre la proporción de valoraciones consistentes y aquellas que se esperarían por azar da el coeficiente de Hambleton y Novick ($p_{H\text{-}N}$):

$$p_{H\text{-}N} = p_c - p_a$$

El cálculo de la proporción de sujetos clasificados de modo consistente se expresa de la manera siguiente:

$$p_c = \sum p_i = \frac{n_{11}}{N} + \frac{n_{00}}{N}$$

Lectura de la fórmula

p_c: Proporción de sujetos clasificados consistentemente en las dos administraciones del test.

N: Número total de sujetos evaluados.

n_{11}: Número de sujetos clasificados como competentes en las dos administraciones del test.

n_{00}: Número de sujetos clasificados como no competentes en las dos administraciones del test.

A pesar de que la expresión que se proporciona contempla clasificar a los individuos solo en dos categorías (competentes y no competentes), la fórmula se puede aplicar independientemente del número de categorías. En este caso habrá que sumar las proporciones de cada una de las categorías establecidas.

Cuando todos los sujetos son clasificados de manera consistente en las dos administraciones del test, p_c toma el valor máximo de 1. Su valor mínimo viene determinado por la proporción de clasificaciones consistentes que se esperaría por azar (p_a). Para calcular esta proporción de clasificaciones consistentes que se esperarían por azar se utilizan las frecuencias marginales de una tabla de contingencia aplicando la fórmula siguiente:

$$p_a = \sum \frac{n_{.j} \cdot n_{i.}}{N^2}$$

Lectura de la fórmula

$n_{.j}$: Número de sujetos clasificados como competentes (o no competentes) por el test A.

$n_{i.}$: Número de sujetos clasificados como competentes (o no competentes) por el test B.

Ejemplo

Supongamos que se administran dos tests paralelos de 20 ítems a un grupo de 16 individuos. Para que el individuo sea clasificado en el grupo de sujetos competentes se requiere que conteste correctamente 14 ítems.

En la tabla siguiente se muestran los datos de este ejemplo. En esta tabla se representan las puntuaciones obtenidas por los 16 individuos en los dos tests (A y B) y la clasificación que a partir del punto de corte establece cada uno de estos tests (clasificación test A y clasificación test B).

Tabla 12. Datos de ejemplo. Coeficiente de Hambleton y Novick

Sujeto	Test A	Test B	Clasificación test A	Clasificación test B
1	10	12	No competente	No competente
2	16	18	Competente	Competente
3	19	20	Competente	Competente
4	8	10	No competente	No competente
5	10	12	No competente	No competente
6	15	16	Competente	Competente
7	14	12	Competente	No competente
8	17	18	Competente	Competente
9	18	19	Competente	Competente
10	13	14	No competente	Competente
11	16	17	Competente	Competente
12	19	17	Competente	Competente
13	12	15	No competente	Competente
14	10	12	No competente	No competente

Sujeto	Test A	Test B	Clasificación test A	Clasificación test B
15	16	14	Competente	Competente
16	18	15	Competente	Competente

A partir de la tabla anterior, se construye la tabla de contingencia, que veremos a continuación.

Tabla 13. Tabla de contingencia. Coeficiente de Hambleton y Novick

		Test B		
Test A		Competentes	No competentes	
	Competentes	9 (n_{11})	1	10 (n_{i1})
	No competentes	2	4 (n_{00})	6 (n_{i2})
		11(n_{1j})	5 (n_{2j})	16 (N)

A partir de la tabla de contingencia podemos calcular el coeficiente de Hambleton y Novic:

$$p_c = \frac{9}{16} + \frac{4}{16} = 0{,}81$$

$$p_a = \frac{10 \cdot 11}{16^2} + \frac{6 \cdot 5}{16^2} = 0{,}546 \approx 0{,}55$$

$$p_{H-N} = 0{,}81 - 0{,}55 = 0{,}26$$

Este resultado indica que el uso de los tests permite mejorar un 26% la clasificación de los sujetos de la que se realizaría por azar.

7.2.2. Coeficiente kappa de Cohen

El coeficiente kappa (Cohen, 1960) permite estudiar el nivel de concordancia en las clasificaciones a partir de dos administraciones del test. Posiblemente este sea el coeficiente de consistencia más extensamente utilizado en la literatura. Su fórmula viene dada por la expresión siguiente:

$$k = \frac{p_c - p_a}{1 - p_a}$$

Donde p_c y p_a son, respectivamente, la proporción de sujetos clasificados de manera consistente y la que se esperaría por azar tal como se ha definido antes. Valores cercanos a 1 indican que la consistencia en la clasificación de los sujetos a partir del test es perfecta, mientras que valores cercanos a 0 indican que la consistencia en la clasificación es debida al azar (en este caso la aplicación de los tests no ha mejorado la consistencia que por azar se podría obtener). En

general, los valores del coeficiente kappa que oscilan entre 0,6 y 0,8 se consideran aceptables y aquellos que se sitúan por encima de 0,8 se interpretan como muy buenos (Landis y Koch, 1977).

A partir del uso del error típico de medida (S_e), propuesto también por Cohen (1960), puede obtenerse el intervalo de confianza y valorar su significación estadística. El error típico de medida y el intervalo de confianza vienen definidos a partir de las expresiones siguientes:

$$S_{e(k)} = \sqrt{\frac{p_c(1-p_c)}{N(1-p_a)^2}}$$

$$IC = k \pm Z_{\alpha/2} \cdot S_{e(k)}$$

A partir de los datos expuestos en el ejemplo anterior el resultado del coeficiente kappa sería:

$$k = \frac{0,81-0,55}{1-0,55} = 0,58$$

Este valor indica la consistencia en la clasificación de los sujetos independientemente de la proporción esperada por el azar. A partir del error típico de medida se obtiene el intervalo de confianza con un nivel de confianza del 95%:

$$S_{e(k)} = \sqrt{\frac{0,81 \cdot (1-0,81)}{16 \cdot (1-0,55)^2}} = 0,22$$

$$IC_{95\%} = 0,58 \pm 1,96 \cdot 0,22$$

$$0,15 \leq k \leq 1$$

Hay que señalar que debido al reducido tamaño de muestra el intervalo construido es excesivamente amplio. Si en vez de tener 16 sujetos fueran 200 los que se hubieran evaluado, se conseguiría un intervalo bastante más preciso:

$$S_{e(k)} = \sqrt{\frac{0,81 \cdot (1-0,81)}{200 \cdot (1-0,55)^2}} = 0,06$$

$$IC_{95\%} = 0,58 \pm 1,96 \cdot 0,06$$

$$0,46 \leq k \leq 0,70$$

7.2.3. Coeficiente de Livingston

Los dos procedimientos anteriores consideran que un error en la clasificación de los individuos es igual de grave, independientemente de que la puntuación del sujeto se sitúe cerca o lejos respecto al punto de corte. No obstante, la lógi-

ca nos dice que si un individuo ha obtenido una puntuación muy distanciada del punto de corte (es decir, muestra una competencia sobradamente alta o muy inferior a la del punto de corte), sería difícil que a partir de una segunda aplicación del test o de un test paralelo el resultado de la clasificación fuera contrario al obtenido en la primera administración. En este sentido, Livingston (1972) propone un nuevo coeficiente para evaluar la fiabilidad de las clasificaciones en el que tiene en cuenta este aspecto. Este nuevo coeficiente se basa en los métodos de pérdida cuadrática que tienen en cuenta la distancia que existe entre el punto de corte y la media de las puntuaciones del test. El coeficiente de Livingston se expresa a partir de la fórmula siguiente:

$$K^2_{xx'} = \frac{r_{xx}S_xS_{x'} + (\bar{x}_x - C)(\bar{x}_{x'} - C)}{\sqrt{[S^2_x + (\bar{x}_x - C)^2]\,[S^2_{x'} + (\bar{x}_{x'} - C)^2]}}$$

Lectura de la fórmula

r_{xx}: Coeficiente de fiabilidad a partir del procedimiento de formas paralelas o test-retests.

S_x y $S_{x'}$: Corresponden, respectivamente, a la desviación típica del test en la primera y segunda administración o en cada una de las formas paralelas del test.

$\bar{x}_x$ y $\bar{x}_{x'}$: Corresponden, respectivamente, a la media del test en la primera y segunda administración o en cada una de las formas paralelas del test.

C: Punto de corte.

S^2_x y $S^2_{x'}$: Corresponden, respectivamente, a la varianza del test en la primera y segunda administración o en cada una de las formas paralelas del test.

A continuación se muestra la aplicación de la fórmula en una situación práctica.

Supongamos que se han diseñado dos formas paralelas de un examen de psicometría. Al aplicarlas a una muestra de estudiantes se obtuvo que la media y desviación típica del test A fue de 5,2 y 2,6, respectivamente. Mientras que en el test B se obtuvo una media de 5,4 y una desviación típica de 2,3. La correlación entre los dos tests fue de 0,83. El punto de corte para determinar el aprobado en la prueba se situó en el 5,5. A partir de estos datos podemos calcular $K^2_{xx'}$:

$$K^2_{xx'} = \frac{0{,}83 \cdot 2{,}6 \cdot 2{,}3 + (5{,}2 - 5{,}5) \cdot (5{,}4 - 5{,}5)}{\sqrt{[2{,}6^2 + (5{,}2 - 5{,}5)^2] \cdot [2{,}3^2 + (5{,}4 - 5{,}5)^2]}} = 0{,}83$$

Este resultado muestra que la concordancia de clasificación a partir de los dos tests es buena.

7.3. Índices de acuerdo que requieren una única aplicación del test

La desventaja principal de los indicadores presentados en el apartado anterior es que hay que aplicar el test dos veces o generar una forma paralela del test. A raíz de este hecho surgió la necesidad de encontrar algún procedimiento en el que solo se necesitará una única administración del test.

La primera propuesta para evaluar la fiabilidad de los TRC que solo requiriera una única administración del test vino de la mano de Livingston (1972), quien propuso una leve modificación a la formulación que se ha presentado en el apartado anterior. A partir de esta primera propuesta siguieron otras muchas[3]. La mayoría de estos procedimientos basan el cálculo de la consistencia de la clasificación utilizando modelos estadísticos que estiman la puntuación que se habría obtenido en una segunda administración del test a partir de la puntuación empírica obtenida por los sujetos. No obstante, hay que decir que la mayoría de estos procedimientos son complejos y poco utilizados en la prác-

[3] Subkoviak (1976), Huynh (1976), Brennan y Kanne (1977), Breyer y Lewis (1994), Livingston y Lewis (1995), Brennan y Wan (2004), Lee (2005), Lee, Brennan, Wan (2009), entre otros.

tica profesional. Por ello, en este apartado solo se abordará la propuesta inicial de Livingston (1972) por ser una de las más sencillas de cálculo y posiblemente la más utilizada en la práctica (Shrock y Coscarelli, 2007).

7.3.1. Coeficiente de Livingston (una única aplicación)

La propuesta que se ha presentado del coeficiente de Livingston en el apartado anterior puede ser fácilmente modificada para que sea válida cuando solo se cuenta con una única administración del test. En este caso, se debe tener en cuenta como coeficiente de fiabilidad el coeficiente de consistencia interna del test. La fórmula propuesta por Livingston (1972) viene dada por la siguiente expresión:

$$K^2 = \frac{r_{xx} S_x^2 + (\bar{x} - C)^2}{S_x^2 + (\bar{x} - C)^2}$$

Lectura de la fórmula

r_{xx}: Coeficiente de fiabilidad a partir del procedimiento de las dos mitades, KR_{20} o alfa de Cronbach.

S_x^2: Varianza de las puntuaciones del test.

$\bar{x}$: Media de las puntuaciones del test.

C: Punto de corte.

Si solo contáramos con los resultados del primer examen de psicometría del ejemplo anterior y el coeficiente alfa de Cronbach fuera de 0,78, K^2 sería:

$$K^2 = \frac{0{,}78 \cdot 2{,}6^2 + (5{,}2 - 5{,}5)^2}{2{,}6^2 + (5{,}2 - 5{,}5)^2} = 0{,}783$$

Hay que tener presente que cuando la media coincide con el punto de corte, K^2 será igual al coeficiente de fiabilidad. Por ejemplo, en el caso anterior se puede observar que la distancia entre la media y el punto de corte no es muy elevada, por lo que el valor de K^2 es muy similar al valor del coeficiente alfa de Cronbach. En cambio, se observa que cuando el punto de corte se distancia de la media, K^2 aumenta. Si ahora consideráramos que el punto de corte es de 6,5, el resultado sería:

$$K^2 = \frac{0{,}78 \cdot 2{,}6^2 + (5{,}2 - 6{,}5)^2}{2{,}6^2 + (5{,}2 - 6{,}5)^2} = 0{,}82$$

Asimismo, se observa que cuando aumenta el coeficiente de fiabilidad del test también aumenta K^2. Ahora, si aumentásemos el coeficiente de alfa de Cronbach a 0,85, observaríamos que también aumentaría el resultado de K^2:

$$K^2 = \frac{0{,}85 \cdot 2{,}6^2 + (5{,}2 - 5{,}5)^2}{2{,}6^2 + (5{,}2 - 5{,}5)^2} = 0{,}852$$

Por ello se demuestra que K^2 siempre será igual o mayor que el coeficiente de fiabilidad del test, y que cuando el coeficiente de fiabilidad sea 1, K^2 tomará también el valor máximo de 1.

$$K^2 \geq r_{xx}$$

7.4. Fiabilidad interobservadores

Ved también

Recordad que hemos visto el coeficiente kappa en el apartado "Coeficiente kappa de Cohen" en este mismo módulo.

Los procedimientos presentados en los apartados previos hacían referencia a la capacidad del test para poder clasificar de manera consistente a los sujetos. En este apartado se presentarán algunos de los indicadores más frecuentes para poder evaluar la consistencia de las evaluaciones realizadas por diferentes jueces. En algunos contextos educativos y de empresa, es frecuente que la competencia de un individuo pueda ser evaluada por diferentes observadores. En estos casos la calidad de la medida depende de la consistencia que se observa entre la evaluación de los observadores. El coeficiente kappa para datos nominales u ordinales y el coeficiente de concordancia para variables continuas son los dos coeficientes más ampliamente utilizados de correlación que se presentarán en este apartado.

7.4.1. Coeficiente kappa

En este caso el coeficiente kappa se aplica cuando se quiere estudiar la concordancia que existe entre las valoraciones realizadas por dos evaluadores. La fórmula, tal como se ha presentado antes, viene dada por la expresión siguiente:

$$k = \frac{p_c - p_a}{1 - p_a}$$

Donde p_c corresponde a la proporción de sujetos clasificados de manera consistente por los dos (o más) evaluadores y p_a corresponde a la proporción de concordancias que se esperaría que sucedieran entre los dos evaluadores por azar.

Supongamos que dos evaluadores han clasificado en dos categorías a 80 trabajadores que han realizado un curso de formación: aquellos que han logrado las competencias básicas para poder desarrollar las nuevas tareas (competentes) y aquellos que se considera que no (no competentes). En la tabla de contingencia siguiente se presentan los resultados de sus valoraciones:

Tabla 14

		Observador 1		
Observador 2		Competentes	No competentes	
	Competentes	49 (n_{11})	11	60 (n_{i1})
	No competentes	1	19 (n_{00})	20 (n_{i2})
		50 (n_{1j})	30 (n_{2j})	80 (N)

$$p_c = \frac{n_{11}}{N} + \frac{n_{00}}{N} = \frac{49}{80} + \frac{19}{80} = 0,85$$

$$p_a = \sum \frac{n_{.j} \cdot n_{i.}}{N^2} = \frac{60 \cdot 50}{80^2} + \frac{30 \cdot 20}{80^2} = 0,523$$

$$k = \frac{p_c - p_a}{1 - p_a} = \frac{0,85 - 0,523}{1 - 0,523} = 0,69$$

Este valor se interpretaría como una consistencia aceptable entre los dos evaluadores.

7.4.2. Coeficiente de concordancia

Lin (1989) propuso un coeficiente de fiabilidad para calcular el grado de acuerdo en las valoraciones realizadas por dos evaluadores cuando estas fueran de carácter continuo. A pesar de que en apartados previos se ha visto que cuando las variables son continuas se aplica el coeficiente de correlación de Pearson (r_{xy}), en este caso no sería adecuado aplicarlo. La razón es que el coeficiente de correlación de Pearson valora solo si el orden de las valoraciones de los dos evaluadores coinciden, pero no si los valores asignados a estas valoraciones son realmente los mismos. Por ello, se podrían obtener coeficientes de correlación de Pearson altos (si la ordenación entre los dos evaluadores es similar), a pesar de que ninguna valoración fuera coincidente. El coeficiente de concordancia propuesto por Lin (1989) permite valorar el grado en el que los valores absolutos otorgados por cada evaluador son concordantes. El coeficiente viene definido por la expresión siguiente:

$$CC = \frac{2 r_{xy} S_x S_y}{S_x^2 + S_y^2 + (\bar{x} - \bar{y})^2}$$

Lectura de la fórmula

r_{xy}: Valor del coeficiente de correlación de Pearson entre las dos valoraciones.

S_x y S_y: Corresponden respectivamente a la desviación típica del evaluador 1 y del evaluador 2.

S_x^2 y S_y^2: Corresponden respectivamente a la varianza del evaluador 1 y del evaluador 2.

$\bar{x}$ e $\bar{y}$: Corresponden respectivamente a la media de las evaluaciones realizadas por el evaluador 1 y 2.

El coeficiente puede tomar valores entre 1 y –1, pero por la cuestión que interesa valorar (coincidencia de las puntuaciones entre los evaluadores) solo tendrá sentido cuando este tome valores positivos. Por otro lado, hay que tener en cuenta que a la hora de valorar el resultado, este coeficiente interpreta el grado de acuerdo de manera mucho más exigente: valores superiores a 0,99 se interpretan como una concordancia casi perfecta; entre 0,95 y 0,99 se habla de una concordancia sustancial; entre 0,90 y 0,95, moderada, y por debajo de 0,90 se considera que la concordancia es pobre.

Ejemplo

Dos evaluadores corrigen de manera independiente los trabajos de 40 alumnos. La correlación de Pearson entre los dos evaluadores es de 0,96. La media de las calificaciones del evaluador A es de 7,23, con una desviación típica de 3,44, mientras que para el evaluador B la media es de 6,33, con desviación típica de 2,88. ¿Cuál es el coeficiente de concordancia entre los dos evaluadores?

$$CC = \frac{2 \cdot 0,96 \cdot 3,44 \cdot 2,88}{3,44^2 + 2,88^2 + (7,23 - 6,33)^2} = 0,91$$

El coeficiente de concordancia indica que el grado de acuerdo entre los dos evaluadores es moderado.

8. Estimación de los puntos de corte

Los procedimientos que se han presentado hasta ahora requieren previamente establecer un punto de corte para calcular la fiabilidad. Tal como apunta Berk (1986), el punto de corte es el punto que permite tomar decisiones y clasificar a los sujetos como competentes (aquellos que dominan el criterio) y no competentes (aquellos que no lo dominan). Dado que el punto de corte es una puntuación empírica del test, tiene asociado un error como puntuación del test que es y un error como puntuación a partir de la cual se toman decisiones. El primer error es el error de medida, que se ha presentado en los apartados previos y que se expresa a partir del error típico de medida. El segundo error es un error de clasificación y puede tomar dos formatos:

- Un individuo que es competente y que erróneamente es clasificado como no competente (falso no competente).

- Un individuo que no es competente y que erróneamente es clasificado como competente (falso competente).

A pesar de que es importante evitar estos dos tipos de errores, hay situaciones en las que cometer un tipo de error u otro tiene repercusiones diferentes y representa una mayor o menor gravedad.

Ejemplo

Supongamos que después de un curso de formación en una empresa se evalúa la competencia de los trabajadores para poder llevar a cabo una nueva tarea. Una vez finalizado el curso, los participantes realizan un examen para valorar si han logrado las competencias básicas para enfrentarse a la nueva tarea. La empresa decide que aquellos que no han superado la prueba participen en unas nuevas charlas para intentar lograr esta competencia. En una situación de este tipo se considera que las consecuencias de clasificar a un sujeto como competente cuando no lo es resultan más graves que clasificar a un sujeto como incompetente cuando en realidad no lo es. En el primer caso, el trabajador no sabrá enfrentarse a la nueva tarea, se equivocará al realizarla, con la posibilidad de que su mala praxis provoque otras consecuencias más graves. En el segundo caso se envía a un sujeto competente a recibir unas charlas adicionales, que seguramente le permitirán lograr con más firmeza los conocimientos que ya tenía adquiridos.

En este apartado abordaremos algunos de los métodos más frecuentes para establecer este punto de corte. Tal como sugieren Cizek y Bunch (2007), cada uno de los métodos mezcla una parte de arte y una parte de ciencia. Por otro lado, hay que tener en cuenta que cada uno de los métodos puede llevar a un resultado diferente (diferente punto de corte y diferente porcentaje de sujetos clasificados en cada grupo) y que en definitiva, a pesar de que algunos métodos son más adecuados en función del tipo de tests o circunstancia, ninguno ha demostrado ser superior a los otros. Todos tienen en común que unos jueces expertos determinarán el punto de corte de manera sistemática a partir de la evidencia que tienen sobre aquello que pretenden valorar. Según Jaeger (1989), los diferentes métodos pueden ser clasificados en dos grandes grupos: por un

lado, los métodos que se centran en la valoración del test (*test centered*) y, por otro, aquellos que se centran en la ejecución de los examinandos (*examinee centered*). No obstante, por motivos didácticos la exposición que se seguirá de los diferentes métodos se realizará según tres criterios:

- Métodos que se basan en la valoración que un grupo de expertos o jueces realizan sobre los ítems de un test (*test centered*). Se presentarán los métodos de Nedelsky (1954), Angoff (1971, 1984) y Sireci, Hambleton y Pitoniak (2004).

- Métodos que se basan en la valoración de un grupo de expertos sobre la competencia de los sujetos (*examinee centered*). Aquí se expondrán los métodos del grupo de contraste (Berk, 1976) y del grupo límite (Zieky y Livingston, 1977).

- Métodos que se basan en la posición del sujeto respecto al grupo normativo, o también llamados métodos de compromiso, que intentan ligar aspectos referentes a los TRC y a los TRN. En este apartado se presentarán los métodos de Hofstee (1983) y Beuk (1984).

8.1. Métodos basados en la evaluación de expertos sobre los ítems

Los métodos que se presentan a continuación tienen en común que los participantes serán jueces expertos sobre el contenido de la materia que evalúan. En todos los casos se les requerirá que valoren si los ítems del test pueden ser contestados correctamente por determinados individuos, sin que por ello sea necesario tener información real sobre la competencia de los individuos a los que se administrará la prueba. De este último aspecto se desprende una de las ventajas principales de estos métodos, y es que pueden ser aplicados antes de que los sujetos contesten el test, dado que no se necesitan datos sobre su ejecución.

8.1.1. Método de Nedelsky

El método que propuso Nedelsky (1954) se aplica sobre ítems de respuesta múltiple. Se utiliza todavía hoy en día de manera amplia en el ámbito académico para determinar el punto de corte y poder determinar si los individuos tienen los conocimientos mínimos sobre una materia concreta. Un concepto clave de este procedimiento y otros que se tratarán en este apartado es que los expertos o jueces deben tener en mente un hipotético grupo de sujetos que se encontrarían en el límite para considerarlos competentes (o incompetentes).

El método se basa en que un grupo de expertos, sobre el contenido que se pretende valorar, determina para cada ítem las alternativas de respuesta que un sujeto con los conocimientos mínimos requeridos sobre la materia para superar la prueba (con la competencia mínima requerida) rechazaría como in-

correctas. El recíproco de las alternativas restantes (1/número de alternativas restantes) es lo que se denomina **valor Nedelsky**. Este valor se interpreta como la probabilidad de que un individuo con una capacidad mínima sobre la materia seleccione la alternativa correcta.

Ejemplo

Para ilustrar el procedimiento, imaginemos un ítem con cinco alternativas de respuesta. El juez determina que dos de las alternativas de respuesta pueden ser descartadas fácilmente por un sujeto que presente los conocimientos mínimos requeridos sobre la materia. El sujeto, por lo tanto, se enfrenta a tres alternativas restantes. El valor Nedelsky en este caso será 1/3 = 0,33. Este procedimiento habría que repetirlo para cada ítem del test, y posteriormente sumar todos los valores obtenidos para cada uno de los ítems. La suma de todos los valores se utiliza como punto de corte del test, que servirá para clasificar a los individuos en las categorías de competentes y no competentes. No obstante, hay que tener en cuenta que es habitual que en el proceso participe más de un juez o evaluador, por lo que habrá que hacer la media de las valoraciones de cada uno de los jueces y posteriormente hacer el sumatorio de las valoraciones medias. En la tabla siguiente se muestra un ejemplo de cálculo del punto de corte a partir de la valoración de 4 jueces a un examen de 12 ítems con 5 alternativas de respuesta.

Tabla 15. Ejemplo de cálculo del método Nedelsky

	Evaluadores				
	A	**B**	**C**	**D**	
Ítems	**Valores Nedelsky**				**Medias**
1	0,5	0,33	0,5	0,5	0,4575
2	1	0,5	1	1	0,875
3	0,2	0,25	0,2	0,25	0,225
4	0,33	0,5	0,25	0,25	0,3325
5	0,33	0,33	0,33	0,25	0,31
6	0,2	0,33	0,5	0,5	0,3825
7	1	0,33	0,5	0,33	0,54
8	0,5	0,25	0,33	0,5	0,395
9	0,25	0,2	0,2	0,2	0,2125
10	0,33	0,33	0,5	0,5	0,415
11	1	1	0,5	0,5	0,75
12	0,33	0,5	0,5	0,5	0,4575
Sumatorio					**5,3525**

En este ejemplo el sumatorio de todas las medias es de 5,35, lo que indica que un sujeto con la competencia mínima requerida sobre la materia para superar la prueba debería contestar 5 ítems correctamente. Así pues, se recomendaría que los sujetos con 5 ítems correctos sean clasificados como competentes, es decir, que aprueben el examen.

La principal limitación del método es que los valores que pueden adquirir las probabilidades asignadas a los ítems son muy limitados y no presentan intervalos iguales entre ellos. Por ejemplo, en el caso anterior los valores de Nedelsky pueden ser: 0,2, 0,25, 0,33, 0,5 y 1, y como se puede observar las distancias entre ellos no son iguales. Este hecho lleva a que, en los casos en los que se observa mayor distancia entre los valores de la probabilidad (por ejemplo entre los valores 0,5 y 1), muchos evaluadores tiendan a puntuar preferiblemente los ítems con una probabilidad de 0,5 antes que con una probabilidad de 1 (probabilidad de que un sujeto con la capacidad mínima requerida sobre el contenido sea capaz de descartar todas las alternativas incorrectas). Esta actuación (otorgar una probabilidad de 0,5 en vez de 1) implica un sesgo a la baja en la puntuación de corte de la prueba, lo que provoca que este método proporcione puntos de corte más bajos en comparación a otros métodos.

8.1.2. Método de Angoff

El método de Angoff (1971) es el más utilizado en la práctica y el que presenta más variantes respecto a su planteamiento inicial (Cizek y Bunch, 2007). Como el método de Nedelsky, consiste en que un grupo de evaluadores base sus juicios teniendo en mente un hipotético grupo de individuos que presenta una competencia mínima sobre la materia para poder superar una prueba. Hay que destacar que, a diferencia del método plateado por Nedelsky, Angoff propuso que los jueces se fijaran en el ítem globalmente y no en cada una de las alternativas de respuesta.

La propuesta inicial de Angoff (1971) consistía en que el evaluador debía determinar si un individuo hipotético, con la competencia mínima requerida, sería capaz de contestar el ítem correctamente. Al ítem se le asignaba un 1, si se considera que el sujeto es capaz de contestar correctamente el ítem, y se puntuaba con un cero si se consideraba que esta persona fallaría el ítem. El sumatorio de las puntuaciones asignadas a cada ítem sería igual a su punto de corte de la prueba, lo que permitiría distinguir aquellos sujetos competentes de los que no presentaran una competencia mínima requerida sobre la materia. No obstante, la versión más extendida de este método requiere la que el propio Angoff planteó en un pie de página en su descripción inicial del método. En este caso proponía una variación del método que consistía en que en lugar de pensar en un individuo concreto, los jueces hicieran sus valoraciones teniendo en mente a un grupo de individuos con una competencia mínima para superar la prueba y que valoraran la proporción de sujetos que contestarían el ítem correctamente. El sumatorio de las probabilidades asignadas a cada uno de los ítems representaría el punto de corte de la prueba. El ejemplo siguiente muestra un hipotético caso de esta última opción.

Ejemplo

Se solicita a cuatro evaluadores que indiquen el porcentaje de sujetos que sería capaz de resolver correctamente cada uno de los 12 ítems de un examen de psicometría. Se indica a los evaluadores que hay que centrar sus valoraciones pensando en un posible grupo

de individuos que supuestamente presenta la competencia mínima requerida sobre la materia. En la tabla siguiente se muestran los resultados de este hipotético caso.

Tabla 16. Ejemplo de cálculo del método de Angoff

	Evaluadores			
	A	**B**	**C**	**D**
Ítems	**Porcentaje de sujetos que resuelven correctamente el ítem**			
1	30	40	45	38
2	80	90	75	85
3	85	70	65	70
4	50	45	50	60
5	60	50	45	55
6	80	90	70	75
7	40	45	50	35
8	20	25	30	10
9	90	80	85	90
10	30	40	35	40
11	40	45	50	50
12	60	65	70	75
Media	55,42	57,08	55,83	56,92

El punto de corte en este caso se obtendrá a partir del cálculo de la media de las puntuaciones otorgadas por los cuatro jueces:

$$C = \frac{55{,}42 + 57{,}08 + 55{,}83 + 56{,}92}{4} = 56{,}31\%$$

Este resultado indica que el punto de corte recomendado por este test sería que los sujetos contestaran correctamente un 56,31% de los ítems, es decir, 6,76 ítems (7 ítems) del conjunto de 12 ítems del test.

Una ventaja importante de este método es que el hecho de evaluar globalmente el ítem permite utilizar ítems de otros formatos, además de los ítems de elección múltiple utilizados en el método de Nedelsky. Esto supone que el método puede ser utilizado cuando los ítems de la prueba son en formato abierto (Hambleton y Plake, 1995). Como se ha comentado antes, las variantes del método son muchas. Una muy usada para evitar opiniones divergentes entre los jueces es pedir a los evaluadores que valoren los ítems en diferentes rondas, empleando un procedimiento similar al que se utiliza en un método Delphi.

8.1.3. Método del consenso directo

El método del consenso directo es un método relativamente reciente propuesto por Sireci, Hambleton y Pitoniak (2004). Algunas de las ventajas destacables respecto a los métodos anteriores son que, por un lado, cada juez experto que participa en el proceso puede expresar directamente su opinión sobre cuál debería ser la localización concreta del punto de corte y, por otro, que permite modificar a los jueces sus valoraciones según las opiniones ofrecidas por otros participantes.

El procedimiento que se sigue en este método requiere que los ítems del test se agrupan en secciones. Estas secciones están formadas a partir de subáreas de contenido homogéneas. La función de los jueces consiste en decir cuántos ítems de cada sección podrían ser contestados correctamente por sujetos que tienen una competencia mínima para poder ser considerados competentes. El sumatorio de los ítems en cada subárea que se considera que se contestarán correctamente será igual a la puntuación de corte.

Inmediatamente después de que los jueces han realizado sus valoraciones, se presentan a todo el panel de expertos los resultados de cada uno de los miembros y se procede a discutir las razones de las posibles diferencias. En esta fase se promueve una discusión abierta entre los jueces para que defiendan y razonen el porqué de sus valoraciones. El objetivo es que la discusión facilite el consenso entre los evaluadores que permita una convergencia en el punto de corte. Una vez finalizada la discusión, se les ofrece la oportunidad de volver a valorar cada una de las secciones del test. Por último, una vez recogidos los datos de la segunda valoración, se comentan en una segunda ronda las diferentes puntuaciones para intentar una mayor convergencia en el punto de corte del test.

Ejemplo

Se solicitó a cuatro jueces que valoraran una prueba de psicometría de 60 ítems. La prueba se dividió en cuatro secciones (análisis de los ítems, fiabilidad, validez y transformación de puntuaciones) y cada sección presentaba un número diferente de ítems. En la tabla siguiente se presenta por columnas el número de ítems que cada juez considera que contestará correctamente un sujeto con conocimientos mínimos para superar la prueba. En la última fila se recoge el sumatorio de ítems de cada experto, que representa su punto de corte recomendado para la prueba. En las dos últimas columnas se muestran las medias y desviaciones típicas (*DT*) de estas valoraciones y el porcentaje de ítems que serían contestados correctamente en cada sección.

Tabla 17. Ejemplo hipotético del método del consenso directo

	Evaluadores					
	A	**B**	**C**	**D**		
Secciones	**Número de ítems contestados correctamente**				**Media y DT**	**%**
Análisis de los ítems (14 ítems)	8	7	8	8	7,75 (0,5)	64,58

	Evaluadores					
	A	B	C	D		
Secciones	Número de ítems contestados correctamente				Media y DT	%
Fiabilidad (20 ítems)	14	12	13	10	12,25 (1,71)	68,06
Validez (16 ítems)	10	11	10	11	10,50 (0,58)	65,63
Transformación de puntuaciones (10 ítems)	6	7	8	7	7,00 (0,82)	70
Sumatorio	38	37	39	36	37,5	62,50

Las desviaciones típicas indican la variabilidad de opinión entre los jueces para cada sección. En el ejemplo, los jueces tienen juicios bastante homogéneos al valorar los ítems referentes a validez (a pesar de que los ítems que hayan podido señalar cada uno de los jueces puedan ser en cada caso diferentes), dado que la desviación típica respecto a su media es menor. En cambio, la sección de fiabilidad presenta un mayor desacuerdo entre los jueces, dado que su desviación típica respecto a la media es la más alta en relación con el resto de las secciones. El objetivo de la discusión de estos resultados con el panel de jueces es que se promueva una mayor convergencia en los datos, que ayude a lograr un consenso en el punto de corte que habría que tomar en esta prueba.

8.2. Métodos basados en la evaluación de expertos sobre la competencia de los sujetos

Los métodos que se presentarán en este apartado se caracterizan por que los jueces, además de ser expertos en la materia que evalúan, han de conocer la competencia de los sujetos. Este hecho implica que las valoraciones que realizarán no serán sobre sujetos hipotéticos, sino sobre individuos reales. Básicamente es en este aspecto donde radica la crítica más feroz sobre estos métodos, dado que a la hora de hacer las valoraciones, los evaluadores pueden estar influenciados no solo por los conocimientos o habilidades de los sujetos, sino por otras variables intrínsecas al individuo que poco tienen que ver con la competencia, como por ejemplo el sexo, la personalidad, la raza, el comportamiento, etc. Otros tipos de sesgo que se han observado a la hora de hacer evaluaciones por parte de jueces se pueden consultar en la revisión de Martínez-Arias (2010).

8.2.1. Método del grupo de contraste

El método del grupo de contraste fue propuesto inicialmente por Berk (1976). En este método es necesario que los jueces clasifiquen a los sujetos en dos grupos en función del nivel de competencia que suponen que exhibirán en la materia evaluada. Es necesario pues que los jueces conozcan sobradamente el rendimiento que exhibirán los sujetos que deben clasificar. En un grupo clasifican a los individuos que consideran que serán competentes y en el otro a los que consideran que serán no competentes. Una vez realizada esta clasificación hay que administrar la prueba a los sujetos y puntuarla. Para determinar el

punto de corte, se suelen utilizar diferentes procedimientos: representar gráficamente las puntuaciones de la prueba de cada uno de los grupos, utilizar como punto de corte algún indicador de tendencia central (media o mediana) o elaborar el análisis de regresión logística. No obstante, el uso de este último procedimiento solo se recomienda si las muestras son suficientemente grandes (Cizek y Bunch, 2007).

Posiblemente, determinar el punto de corte a partir de la representación gráfica sea el método más sencillo. Consiste en representar en una misma gráfica las distribuciones de los grupos: los que han sido clasificados como competentes por los jueces y los que han sido valorados como no competentes (figura 2). La intersección entre ambas distribuciones sería el punto de corte de la prueba. Idealmente se busca aquel punto de corte en el que todos los sujetos valorados como competentes se sitúen por encima del punto de corte, mientras que los no competentes se encuentren por debajo. En la práctica, tal como se ve en la figura 2, hay un solapamiento entre las dos distribuciones, con lo que en función de dónde se sitúe el punto de corte habrá más sujetos competentes clasificados que no competentes o al revés.

Por ejemplo, si observamos la figura 2, si se moviera la línea vertical (que indicaría el punto de corte) hacia la izquierda, el número de falsos negativos se reduciría, es decir, se disminuiría la posibilidad de considerar un competente como no competente (a pesar de que se aumentaría el número de sujetos no competentes, que se consideran competentes). Mientras que si la línea se moviera hacia la derecha, disminuirían los falsos positivos, es decir, aquellos que siendo no competentes se los considera competentes. Así pues, para fijar el punto de corte, hará falta en cada caso valorar qué tipo de error se quiere evitar.

Buscar el punto de corte a partir de la figura 2 puede resultar una tarea bastante subjetiva, por lo que algunos autores prefieren basar la posición del punto de corte a partir de la media o la mediana de los dos grupos.

A partir de los datos que hemos representado en la figura 2, a modo de ejemplo, supongamos que un grupo de profesores clasificó a sus alumnos de la asignatura de *Psicometría* en competentes y no competentes (n = 258). Los profesores clasificaron a 90 alumnos como no competentes y a 168 como competentes. Los estudiantes hicieron la prueba y la mediana del grupo de no competentes fue de 22 puntos y la del grupo de competentes de 44. El punto medio entre las dos medianas (en este caso 33) puede funcionar como punto de corte de la prueba.

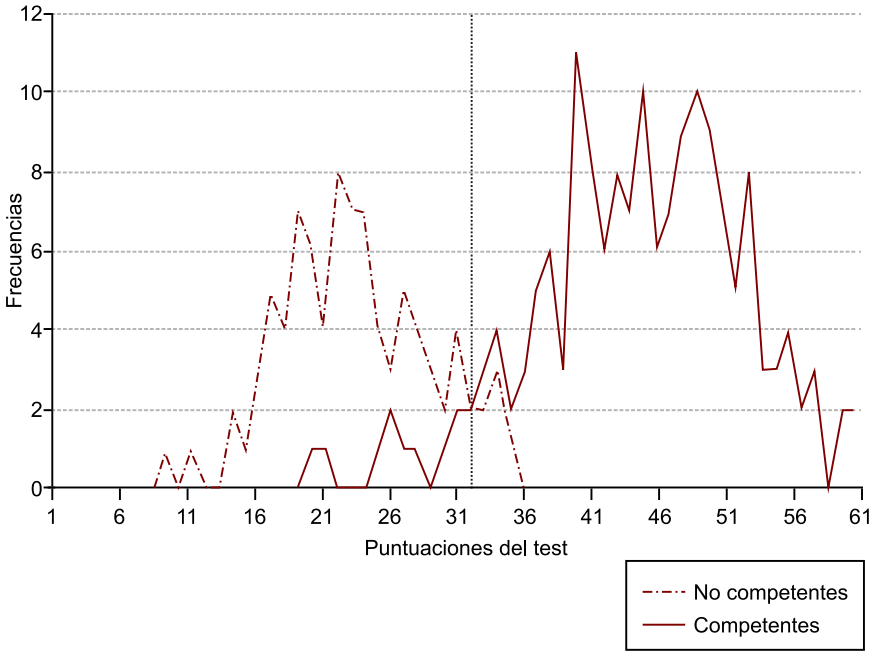

Figura 2. Método del grupo de contraste. Distribuciones del grupo de competentes y no competentes en las puntuaciones de un test

8.2.2. Método del grupo límite

En muchas ocasiones puede resultar difícil a los jueces clasificar a los sujetos claramente en dos grupos, como competentes o no competentes. En este sentido, el método del grupo límite (Zieky y Livingston, 1977) viene a paliar este aspecto y puede ser utilizado como alternativa al método de los grupos de contraste. En este método, generalmente se pide a los jueces que clasifiquen a los sujetos en tres grupos: un grupo de sujetos que claramente son competentes, otro grupo que claramente se percibe como no competentes y finalmente otro que se situaría entre los dos grupos anteriores. Es decir, un grupo límite, en el que los sujetos tendrían una ejecución que los situaría en medio del grupo de competentes y no competentes. Una vez se ha clasificado a los sujetos se administra la prueba. Habitualmente, el valor de la mediana del grupo límite en la prueba es el que se utiliza como punto de corte. Dada la sencillez del método, otros autores (Plake y Hambleton, 2001) propusimos una versión alternativa para valorar grupos de sujetos con competencia básica, notable y avanzada.

8.3. Métodos de compromiso

El término *método de compromiso* fue propuesto por Hofstee (1983) para recoger la idea de que se necesitaban nuevos procedimientos que combinaran, por un lado, los conocimientos que un sujeto presentaba sobre la materia (valoración de la competencia en términos absolutos), pero también el nivel de ejecución que presentaba respecto a su grupo normativo (valoración de la competencia en términos relativos). Los métodos que se han presentado hasta ahora para fijar el punto de corte son métodos que se basan exclusivamente en la competencia que el sujeto exhibe sobre la materia que debe evaluar (valoración en términos absolutos). No obstante, en la práctica los juicios que emiten los

evaluadores no pueden estar estrictamente basados en un criterio absoluto, sino que en cualquier proceso de evaluación se tiene en cuenta información referente al grupo normativo de referencia.

Un ejemplo muy clarificador que ilustra el uso combinado de información sobre el criterio y sobre el grupo normativo en cualquier proceso de evaluación lo proporcionan Cizek y Bunch (2007). Estos autores proponen una situación en la que se valora la competencia de control de esfínteres de un niño. Los padres siguen las rutinas habituales para que su niño adquiera un control adecuado, hasta que llega el gran día en el que pueden afirmar que su hijo ha adquirido la competencia. Hasta aquí todo parece bastante corriente, pero ¿cambiaría la percepción de la situación si dijéramos que el niño de este ejemplo tiene 9 años? Evidentemente, sí.

Este ejemplo demuestra que las personas utilizan múltiples fuentes para valorar las diferentes situaciones a las que se enfrentan y que cualquier intento de anular o reducir alguna de estas fuentes ocasionaría una valoración artificial y poco ajustada a la realidad.

Los métodos que se presentan a continuación aceptan que los tests referidos a la norma y al criterio están, en realidad, bastante unidos a la hora de valorar a los individuos, dado que utilizan las normas para poder fijar el criterio. En este apartado se presentarán dos de los procedimientos más habituales: el método de Hofstee (1983) y el método de Beuk (1984).

8.3.1. Método de Hofstee

Hofstee propuso este método en 1983, cuando después de impartir la misma asignatura durante algunos años se encontró que en aquel curso los alumnos presentaban un rendimiento muy bajo respecto a años anteriores (sin que ningún aspecto significativo respecto al material, profesores, etc., variaran de un año al otro). Después de rebajar la nota de corte a un 4,5, solo el 55% de los estudiantes aprobaban la asignatura (cuando en años anteriores la nota de corte se había situado en un 6 y aun así aprobaba el 90% de los alumnos). El método que Hofstee propone requiere que los evaluadores respondan a cuatro preguntas sobre los sujetos que serán evaluados:

1) ¿Cuál es el punto de corte más alto que se considera aceptable, a pesar de que todo el mundo llegara a esta puntuación? Se simboliza por $k_{máx}$.

2) ¿Cuál es el punto de corte más bajo que se considera aceptable, a pesar de que nadie llegara a esta puntuación? Se simboliza por $k_{mín}$.

3) ¿Cuál es el porcentaje máximo de sujetos que se toleraría que no superaran la prueba? Se simboliza por $f_{máx}$.

4) ¿Cuál es el porcentaje mínimo de sujetos que se toleraría que no superaran la prueba? Se simboliza por $f_{mín}$.

Como se observa, dos de las preguntas se basan en el nivel de conocimiento que los evaluados deben presentar ($k_{máx}$ y $k_{mín}$). Estas dos cuestiones se miden a partir del porcentaje de ítems que hay que contestar correctamente. Las otras dos preguntas se basan en el porcentaje de no competentes que se toleraría

dada una determinada prueba de evaluación ($f_{máx}$ y $f_{mín}$). A partir de estos cuatro puntos y la distribución empírica de la prueba, se encuentra el punto de corte (x) óptimo para este método de compromiso.

En la figura 3 se representa gráficamente cómo encontrar este punto de corte. Por un lado, en el eje de abscisas se representa el porcentaje de ítems correctos de la prueba y, por otro, en el eje de ordenadas el porcentaje de personas que no superan la prueba. En esta gráfica también se representa la distribución empírica de la prueba, que muestra que a medida que aumenta el número de ítems que hay que contestar correctamente (eje de abscisas), aumenta el porcentaje de personas que no superan la prueba. Los valores k y f se representan con dos puntos en el eje de coordenadas y se unen con una línea recta que cruza la distribución empírica de la prueba (en realidad se espera que en la mayoría de las ocasiones atravesará la distribución empírica). El punto en el que se cruza la distribución empírica y la recta k-f indicará el porcentaje de ítems correctos que habrá que exigir a la prueba (el punto de corte) y el correspondiente porcentaje de sujetos que no superan la prueba si se utiliza este punto de corte. Este punto de corte se simboliza en la figura con x.

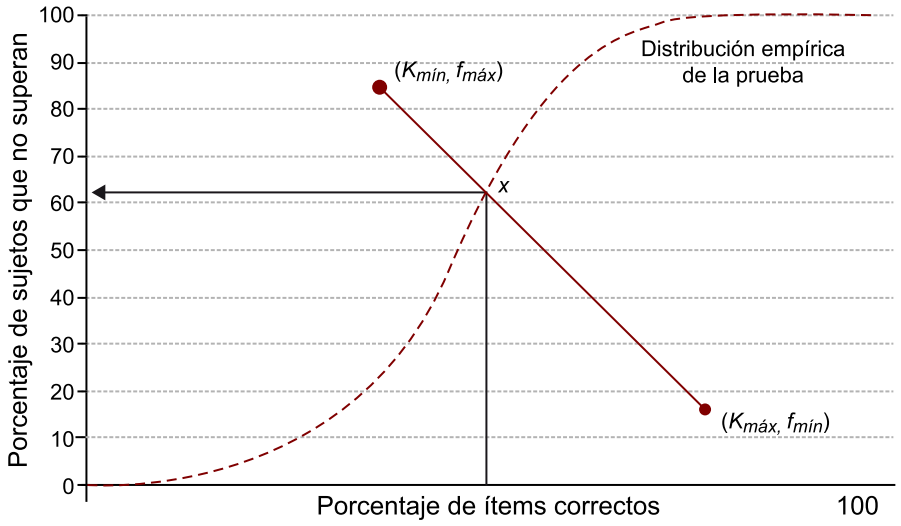

Figura 3. Representación gráfica del método de Hofstee

8.3.2. Método de Beuk

El método propuesto por Beuk (1984) surgió como una simplificación del método de Hofstee (1983). Como el anterior, presupone que los evaluadores tienen una idea más o menos clara sobre cuál es el punto de corte que sería necesario aplicar a la prueba y cuál debería ser el porcentaje de personas que tendrían que superarla. En este caso, se pide a los evaluadores que respondan a las siguientes dos preguntas:

1) ¿Cuál debería ser el porcentaje mínimo de ítems que habría que contestar correctamente para superar la prueba? Este valor se simboliza con x.

2) ¿Cuál es el porcentaje de personas que se espera que superen la prueba? Este valor se simboliza con y.

A partir del cálculo de la media de las valoraciones de los evaluadores a estas dos preguntas ($\overline{x}$ e $\overline{y}$) y la distribución empírica obtenida de la aplicación de la prueba, se obtiene el punto de corte. El ejemplo que se presenta a continuación y la figura 4 muestran cómo obtener el punto de corte a partir de este método.

Supongamos que como media los evaluadores han determinado que haría falta que un 60% de los ítems de la prueba se contestaran correctamente ($\overline{x}$) y que sería necesario que un 70% de las personas que realizaran la prueba la superaran ($\overline{y}$).

En la figura 4 se representan estos dos puntos y su intersección se simboliza con la letra A. Se representa la distribución empírica de la prueba (línea de puntos), que en este caso será decreciente, dado que el número de personas que superarán la prueba (eje de ordenadas) disminuirá a medida que el porcentaje de ítems que hay que contestar de manera correcta (eje de abscisas) aumente.

El paso siguiente es calcular las desviaciones típicas a las dos preguntas formuladas a los evaluadores. El cociente entre las dos desviaciones típicas (S_x/S_y) será la pendiente de la recta que atravesará la distribución empírica de la prueba. El punto en el que la recta atraviesa la distribución empírica se simboliza con el punto B. Finalmente, como se puede ver en la figura 4, la proyección de este punto en el eje de abscisas proporcionará el punto de corte de la prueba (porcentaje de ítems correctos) y su proyección sobre el eje de ordenadas proporcionará el porcentaje de sujetos que superarán la prueba.

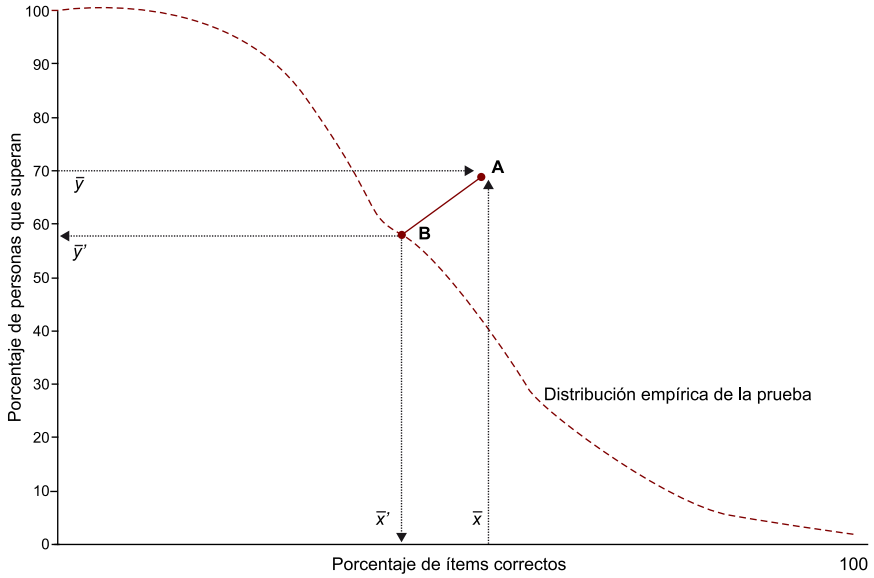

Figura 4. Representación gráfica del método de Beuk

Bibliografía

Angoff, W. H. (1971). Scales norms and equivalent scores. En R. L. Thorndike (Ed.), *Educational measurement* (2.ª ed.). Washington: American Council on Education.

Angoff, W. H. (1984). *Scales norms and equivalent scores*. Princeton: NJ. Educational Testing Service.

Barbero, M. I., Vila, E., y Suárez, L. C. (2003). *Psicometría*. Madrid: UNED.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45 (2), 4-9.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced test. *Review of Educational Research*, 56 (1), 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, 21 (2), 147-152.

Brennan, R. L. y Wan, L. (2004). A bootstrap procedure for estimating decision consistency for single-administration complex assessments. (CASMA Research Report, núm. 7). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Brennan, R., L. y Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14 (3), 277-289.

Breyer, F. J., y Lewis, C. (1994). Pass-fail reliability for tests with cut scores: A simplified method. *ETS Research Report*, 94-39. Princeton, NJ: Educational Testing Service.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37-46.

Cohen, R. J. y Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7.ª ed.). Boston: McGraw-Hill.

Crocker L. y Algina, J. (1986). *Introduction to classical modern test theory.* New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alfa and the internal structures of tests. *Psychometrika*, 16 (3), 297-334.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 30 (3), 357-370.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder -Richardson coefficient twenty is the same for two tests. *Psychometrika*, 34 (3), 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, 45 (1), 99-105.

Flanagan, J. C. (1937). A note on calculating the standard error of measurement and reliability coefficients with the test scoring machine. *Journal of Applied Psychology*, 23 (4), 529.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psichometrika*, 10 (4), 255-282.

Hambleton, R. K. y Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced test. *Journal of Educational Measurement*, 10 (3), 159-170.

Hambleton, R. K. y Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8 (1), 41-55.

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

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

Jaeger, R. M. (1989). Certification of student competence. En R. L. Linn (Ed.), *Educational measurement* (3.ª ed.). New York: Macmillan.

Kapplan, R. M. y Saccuzo, D. P. (2009). *Psychological testing. Principles, applications and issues* (7.ª ed.). Edition. Belmont, CA: Wadsworth.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, 28 (3), 221-238.

Kuder, G. F. y Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2 (3), 151-160.

Landis, J. R. y Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159-74.

Lee, W. (2005). Classification consistency under the compound multinomial model. (CASMA Research Report, núm. 13). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Lee, W., Brennan, R. L., y Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement*, 33 (5), 374-390.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics,* 45 (1), 255-268.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 9 (1), 13-26.

Livingston, S. A. y Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32 (2), 179-197.

Martínez-Arias, R. (2010). La evaluación del desempeño. *Papeles del psicólogo*, 31 (1), 85-96.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide

Murphy, K. R. y Davidshofer, C. O. (2005). *Psychological testing. Principles and applications* (6.ª ed.). Upper Saddle River, NJ: Prentice Hall.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14 (1), 3-19.

Nunnally, J. C. (1978). Psychometric theory (2.ª ed.). New York: McGraw-Hill.

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

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99-103.

Shrock, S. A. y Coscarelli, W. C. (2007). Criterion-referenced test development: Technical and legal guidelines for corporate training and certification (3.ª ed.). San Francisco, CA: John Wiley and Sons.

Sireci, S. G., Hambleton, R. K., y Pitoniak, M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review*, 15 (1), 21-25.

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13 (4), 265-275.

Woodruff, D. J. y Feldt, L. S. (1986). Test for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika*, 51(3),393-413.

Zieky, M. J. y Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service.