3

Modelo Clásico y fiabilidad

Introducción

En las Ciencias clásicas (Medicina, Física, Química,...) existen aparatos, con márgenes de error especificados, para medir características muy diversas como son la tensión arterial, la temperatura o la concentración de determinados elementos químicos. A pesar de la diversidad de atributos que pueden medirse, cada uno de estos instrumentos debe satisfacer siempre cuatro tipos de exigencias:

1. Que la medición sea *fiable* o replicable; es decir, que al repetir las medidas de la misma magnitud se produzcan resultados iguales o parecidos. Por ejemplo, esperaremos obtener medidas similares independientemente de si el termómetro es digital o de mercurio, de la persona que toma la temperatura o, si el intervalo entre medidas es suficientemente corto, del momento en que se realizan.
2. Que las inferencias sobre los atributos que se realizan a partir de las medidas observadas sean *válidas.* Nuestras inferencias serán válidas si son ciertos los principios teóricos en los que se fundamentan. Por ejemplo, a partir del principio físico de la dilatación y teniendo en cuenta el valor del coeficiente de dilatación del mercurio, podemos deducir la temperatura de un objeto a partir de la marca del mercurio en un tubo de cristal.
3. Que se siga el *protocolo de aplicación del instrumento* y que se atienda al mantenimiento de éste, si es necesario. Por ejemplo, para detectar la presencia de fiebre debemos saber en qué parte del cuerpo poner el termómetro y durante cuánto tiempo.
4. Que se tenga en cuenta su *rango de aplicabilidad.* Cualquier instrumento tendrá un rango de aplicabilidad según los niveles de atributo entre los que permite discriminar. En nuestro ejemplo, un termómetro para medir la temperatura corporal no será útil para medir las altas temperaturas en un horno.

Las anteriores exigencias también deberían mantenerse para cualquier instrumento de medición en Psicología y disciplinas afines. Podemos pensar en las consecuencias que tie-

ne para el psicólogo de selección que un test no proporcione una buena información de los niveles de inteligencia de los aspirantes; o las consecuencias que puede tener para un estudiante que se le aplique una prueba de admisión a la universidad de forma inapropiada o incorrecta; también un psicólogo clínico que utiliza un test de depresión en su labor profesional, debe tener un alto grado de certeza de que las puntuaciones que proporciona el test resultan buenas cuantificaciones de los niveles de depresión de sus pacientes; los ejemplos son innumerables…

Sin embargo no resulta fácil medir atributos psicológicos pues no existen modelos teóricos tan desarrollados y consensuados como los de las ciencias clásicas. Aún así, los psicólogos han intentando definir modelos teóricos que permitan inferir constructos teóricamente relevantes (o, al menos, predictivos) como la *depresión* a partir de los comportamientos o las respuestas de los evaluados a una serie de ítems. Los procedimientos para evaluar la verosimilitud de esas inferencias se abordarán en el capítulo sobre validación de las medidas.

El presente capítulo se centra en la primera exigencia (fiabilidad de las medidas) ya que si las medidas no se replican de una situación a otra, difícilmente podremos defender ninguna inferencia o predicción. Si las puntuaciones no se replican diremos que son poco precisas o poco fiables. En este capítulo, revisaremos el modelo matemático propuesto por Spearman (1904a; 1904b; 1907) que permite operativizar el concepto de fiabilidad y valorar las posibles repercusiones de la falta de fiabilidad en nuestras decisiones.

La Teoría Clásica de los Tests

La principal idea del modelo de Spearman es que debemos distinguir entre el valor real del atributo que medimos (la puntuación verdadera) y la medida falible que obtenemos en el proceso de medición (la puntuación observada). Las medidas que tomamos incluyen un cierto grado de error. El error de medida expresa el grado en que nuestra medida se desvía del valor real.

La idea de partida para poder graduar la precisión de un instrumento es que, cuanto más preciso es, más se replicarán nuestras observaciones en sucesivas mediciones. Aunque el planteamiento parece sencillo, dos obstáculos acompañan desde el principio a la medición de cualquier variable psicológica. El primero es que es difícil obtener medidas repetidas independientes de la misma persona con el mismo instrumento en un intervalo corto de tiempo. En Psicología esa falta de independencia (p.ej., porque haya recuerdo de las respuestas dadas en la primera ocasión) puede tener efectos en la replicabilidad de las medidas que no se asocian a la precisión del instrumento. Por tanto, es importante definir bien lo que se entiende por replicabilidad. El segundo obstáculo es la imposibilidad de obtener mediciones directas. No podemos tener acceso directo al valor real de un atributo psicológico. Por tanto, será importante definir bien qué se entiende por puntuación verdadera y por error.

Spearman desarrolló un modelo formal denominado *Modelo Clásico* o *Modelo Lineal Clásico*, fundamentado en diversos supuestos a partir de los cuales se definen los conceptos de puntuación verdadera y error y se extraen determinadas consecuencias de aplicabilidad práctica para cuantificar el tamaño de esos errores y corregir su efecto. Cuando aplicamos un test pretendemos que sea preciso, es decir, que la variabilidad de los evalua-

dos según su puntuación en el test refleje su variabilidad real en el atributo. Pues bien, el Modelo Clásico nos permite deducir, de forma elegante, qué parte de la variabilidad en las puntuaciones en un test se debe a la variabilidad en el atributo medido y qué parte se explica por la presencia de errores en el proceso de medición.

A las ideas de Spearman se han sumado posteriormente las aportaciones de otros muchos investigadores. El armazón teórico del Modelo Clásico se conoce como *Teoría Clásica de los Tests (TCT)* y se trata del principal modelo de referencia para la construcción y evaluación de tests psicológicos. En castellano, información sobre el Modelo Clásico puede encontrarse en varios manuales (p.ej., Martinez Arias, 1995; Martinez Arias, Hernandez-Lloreda y Hernandez-Lloreda, 2006; Muñiz, 1998; Nunnally y Bernstein, 1995). En inglés, existen numerosos manuales sobre la Teoría de los Tests (p.ej., Allen y Yen, 1979; Crocker y Algina, 1986; Gulliksen, 1950; De Gruijter y van der Kamp, 2003; Furr y Bacharach, 2008).

Los supuestos fundamentales del Modelo Clásico

El Modelo Clásico se sustenta en varios supuestos muy simples (Feldt y Brennan, 1989; Haertel, 2006). Considere que para medir el atributo psicológico disponemos de varios tests distintos a los que llamaremos *formas*. Por ejemplo, si quisiéramos medir la Depresión podríamos tener un banco de ítems enorme con muchas de las preguntas posibles. A partir de esas preguntas, podríamos construir distintos tests con especificaciones idénticas (igual número de ítems, contenido similar, etc.). Cada uno de esos tests sería una forma del test. A las puntuaciones que obtienen los evaluados en las distintas formas las denominaremos como variables X_1, X_2,…, X_f; a continuación se describe qué propiedades deben tener esas formas para que podamos estimar la precisión de cualquiera de ellas.

Primer supuesto: el modelo lineal

El primer supuesto establece que la puntuación observada de una persona i en una forma f de un test (X_{if}) se descompone linealmente en dos componentes hipotéticos, la puntuación verdadera de la persona (V_i), que es una constante para cada persona i, y el error de medida que se comete al medir el rasgo con el test f (E_{if}):

$$X_{if} = V_i + E_{if} \qquad [3.1]$$

La puntuación verdadera refleja por tanto la puntuación en el atributo tal y como lo mide un test con esas especificaciones; esto quiere decir que las puntuaciones verdaderas de una persona en dos tests con distintas especificaciones, por ejemplo distinto número de ítems, no serán iguales. Observe que la puntuación V_i no lleva el subíndice f; se asume que la puntuación verdadera del evaluado i es la misma en cada una de las formas:

$$V_i = V_{i1} = V_{i2} = \cdots = V_{if}$$

El error de medida depende de diferentes factores (propios de la persona, del test y de la situación) que hacen que su puntuación empírica, X, no sea exactamente su nivel de atributo, V. Por ejemplo, en una prueba de conocimientos pueden constituir fuentes de error el nivel de ansiedad, la falta de motivación para responder, el ruido en el aula, la adecuación de las instrucciones de aplicación, el nivel de riesgo asumido por el que responde, la suerte que se tiene al responder a las preguntas de las que no se sabe la respuesta, qué preguntas concretas aparecen en la prueba, etc.

Por tanto, el *error de medida* se establece como la diferencia entre la puntuación empírica y la verdadera:

$$E_{if} = X_{if} - V_i \qquad [3.2]$$

Considerando todo lo anterior, el Modelo Clásico puede expresarse en términos de variables como:

$$X_f = V + E_f \qquad [3.3]$$

Para que se comprenda lo que significa cada uno de los términos, obsérvese la estructura de la siguiente matriz de datos en la Tabla 3.1. V, E_f (E_1, E_2,…) X_f, (X_1, X_2,…) son las variables (el subíndice f indica la forma aplicada) y V_i, E_{if}, X_{if}, indican los valores de las variables para el i-ésimo evaluado.

Tabla 3.1. Estructura de una matriz de datos si aplicáramos distintas formas paralelas a un grupo de evaluados y fueran conocidas las puntuaciones verdaderas (V)

	Puntuación verdadera	*Error con el test 1*	*Puntuación empírica en el test 1*	*Error con el test 2*	*Puntuación empírica en el test 2*	*Error con el test 3*	*Puntuación empírica en el test 3*	…
	V	E_1	$X_1 = V + E_1$	E_2	$X_2 = V + E_2$	E_3	$X_3 = V + E_3$	…
Evaluado 1	V_1	E_{11}	X_{11}	E_{12}	X_{12}	E_{13}	X_{13}	…
Evaluado 2	V_2	E_{21}	X_{21}	E_{22}	X_{22}	E_{23}	X_{23}	…
Evaluado 3	V_3	E_{31}	X_{31}	E_{32}	X_{32}	E_{33}	X_{33}	…
Evaluado 4	V_4	E_{41}	X_{41}	E_{42}	X_{42}	E_{43}	X_{43}	…
Evaluado 5	V_5	E_{51}	X_{51}	E_{52}	X_{52}	E_{53}	X_{53}	…
…	…	…	…	…	…	…	...	…

Ejemplo 3.1. El Modelo Clásico lineal

Supongamos que, para un grupo de personas, conocemos las puntuaciones V, E_f y X_f, en múltiples formas del test (en realidad, sólo podemos conocer las puntuaciones X; las restantes puntuaciones se proponen únicamente por razones didácticas):

Tabla 3.2. Puntuaciones verdaderas (*V*), puntuaciones observadas (*X*) y errores de medida (*E*) al aplicar varias formas de un test[1,2]

Formas		*Forma* 1		*Forma* 2		*Forma* 3		…
	V	E_1	X_1	E_2	X_2	E_3	X_3	…
Evaluado 1	12	–2	10	0	12	0	12	…
Evaluado 2	11	0	11	–2	9	–2	9	…
Evaluado 3	11	0	11	2	13	2	13	…
Evaluado 4	12	2	14	0	12	0	12	…
Evaluado 5	4	0	4	0	4	0	4	…
…	…	…	…	…	…	…	…	

Puede observarse que la puntuación empírica del tercer evaluado en el segundo test (X_{32}) es 13 por lo que se sobrestima su puntuación verdadera (V_3), que es 11, en 2 puntos (que es el error, E_{32}). Los evaluados 2 y 3 tienen la misma puntuación verdadera (11); sin embargo sus puntuaciones observadas cuando se aplica la forma 2 del test son distintas (9 y 13), lo que expresa que se comete cierto error de medida (subestimación en el primer caso y sobreestimación en el segundo).

Segundo supuesto

El problema es que *E* y *V* son desconocidas. Sin embargo, podemos obtener información sobre ellas si se plantean determinados supuestos adicionales. En muchos contextos parece razonable asumir que los errores serán unas veces positivos (por sobreestimación de la puntuación verdadera) y otras veces negativos (por subestimación). Por ejemplo, es posible que al responder a un examen la persona reciba más preguntas de los temas que más ha estudiado; en este caso, su puntuación empírica será una sobreestimación de lo que sabe (error positivo). En otros exámenes sucederá lo contrario (error negativo). Por tanto, desde un punto de vista conceptual, la mejor estimación del verdadero conocimiento del evaluado *i* será el promedio (valor esperado) de las puntuaciones empíricas que obtendría en un número elevado de aplicaciones. Este es el segundo supuesto:

$$V_i = \varepsilon_f (X_{if}) \qquad [3.4]$$

[1] En los ejemplos que siguen se consideran que *X*, *V* y *E* son puntuaciones discretas. Sin embargo, este no es un requerimiento del Modelo Clásico y, de hecho, las estimaciones de *V* y *E* pueden ser números decimales.

[2] En los ejemplos que siguen se muestran tablas incompletas de datos. Por ejemplo, en la Tabla 3.2 se muestran los datos de 5 evaluados seleccionados de una población más amplia. El hecho de que el número de evaluados es más amplio se indica mediante puntos suspensivos. Por tanto, cuando se informe del resultado de cualquier cálculo realizado con la población total (sumas, medias y desviaciones típicas, etc.) este no coincidirá generalmente con el que se obtiene a partir de los datos de los 5 evaluados. Lo mismo puede decirse con respecto al número de formas aplicadas: aunque se muestran los resultados obtenidos en 3 formas se asume que se han aplicado muchas más.

Donde el símbolo $\varepsilon_f(\cdot)$ indica valor esperado de la variable dentro del paréntesis a través de *f*. Otra forma de presentar el segundo supuesto es:

$$\varepsilon_f(E_{if}) = 0 \qquad [3.5]$$

que es equivalente a decir que los errores que cometemos no son sistemáticos (el valor esperado de los errores a través de un conjunto de mediciones independientes de la misma persona es 0). Resulta fácil comprobar la igualdad entre [3.4] y [3.5], puesto que:

$$\varepsilon_f(E_{if}) = \varepsilon_f(X_{if} - V_i) = \varepsilon_f(X_{if}) - \varepsilon_f(V_i) = \varepsilon_f(X_{if}) - V_i = V_i - V_i = 0$$

Además, se asume que el valor esperado del error de medida es igual a 0, no sólo para cualquier persona, sino también para un grupo de evaluados a los que se aplica una única forma *f*:

$$\varepsilon_i(E_{if}) = \mu_{E_f} = 0 \qquad [3.6]$$

Veamos en el siguiente ejemplo lo que implica el segundo supuesto.

Ejemplo 3.2. Segundo supuesto

Tabla 3.3. Media de *X* y *E* para cada evaluado a través de las distintas formas y para cada forma a través de los distintos evaluados

		Forma 1		*Forma* 2		*Forma* 3			*Media de X (a través de las formas)*	*Media de E (a través de las formas)*
	V	E_1	X_1	E_2	X_2	E_3	X_3		$\varepsilon_f(X_{if})$	$\varepsilon_f(E_{if})$
Evaluado 1	12	–2	10	0	12	0	12	…	12	0
Evaluado 2	11	0	11	–2	9	–2	9	…	11	0
Evaluado 3	11	0	11	2	13	2	13	…	11	0
Evaluado 4	12	2	14	0	12	0	12	…	12	0
Evaluado 5	4	0	4	0	4	0	4	…	4	0
…	…	…	…	…	…	…	…		…	…
μ_{E_f}		0		0		0				

La media de las puntuaciones empíricas para el evaluado 2 coincidiría justamente con su puntuación verdadera (11). Es decir, la puntuación 11 expresa su nivel promedio en puntuaciones empíricas de depresión a través de las distintas aplicaciones (11, 9, 9,…). Ese promedio puede considerarse la mejor estimación de su puntuación verdadera. Por otro lado, el segundo supuesto implicaría que las medias de los errores para un evaluado a través de distintas formas y para una forma a través de distintos evaluados son cero. Por ejemplo, para el segundo evaluado la media de los errores (0 , –2, –2, …) sería 0. Tam-

bién, según el Modelo Clásico, la media de los errores en la forma 1 (–2, 0, 0, 2,…) sería 0.

Tercer, cuarto y quinto supuestos

Si el error cometido al utilizar una forma no es sistemático parece razonable asumir que los errores en una forma (E_f) no correlacionan con las puntuaciones verdaderas (V), ni con los errores en otra forma ($E_{f'}$) ni con las puntuaciones verdaderas en otro test (V_k):

Tercer supuesto:

$$\rho_{E_f V} = 0 \qquad [3.7]$$

Cuarto supuesto:

$$\rho_{E_f E_{f'}} = 0 \qquad [3.8]$$

Quinto supuesto:

$$\rho_{E_f V_k} = 0 \qquad [3.9]$$

Eso supone asumir, por ejemplo, que las personas que tienen errores elevados (bajos) no tienen porqué tener asociados puntuaciones verdaderas elevadas (bajas) en ese test (u otro) ni errores elevados (bajos) en otras formas del test.

Descomposición de la varianza de las puntuaciones empíricas en un test

Hasta ahora hemos observado una serie de supuestos sobre los errores de medida. Como ya hemos insistido, en la aplicación real de un test sólo se conocen las puntuaciones X de las personas, por lo que los supuestos planteados no pueden, en general, someterse a contrastación empírica. ¿Para qué sirven entonces estos cinco supuestos? ¿Qué nos dicen sobre las puntuaciones de las personas que responden a un test? Si asumimos que los supuestos son lógicos y razonables, podemos obtener indicadores que nos proporcionen información sobre el tamaño de los errores cometidos con un test.

Para empezar, si aceptamos los supuestos, podremos delimitar algunas de las características de las distribuciones de las variables implicadas en la población:

$$\mu_{X_f} = \mu_V \qquad [3.10]$$

$$\sigma^2_{X_f} = \sigma^2_V + \sigma^2_{E_f} \qquad [3.11]$$

Es decir, la media de las puntuaciones observadas en un test *f* coincidirá con la media de las puntuaciones verdaderas y, lo más importante, la varianza de las puntuaciones observadas en un test *f* se puede descomponer en varianza de las puntuaciones verdaderas y varianza de los errores.

En efecto, si tenemos una variable X_f que es combinación lineal de otras variables V y E_f, tal que:

$$X_f = V + E_f$$

puede demostrarse que la media y varianza de la variable X se deriva de las medias y varianzas de las variables V y E_f; esto es:

$$\mu_{X_f} = \mu_V + \mu_{E_f}$$

y dado [3.6] se deriva [3.10]. Además:

$$\sigma^2_{X_f} = \sigma^2_V + \sigma^2_{E_f} + \rho_{VE_f}\sigma_V\sigma_{E_f}$$

y dado [3.7] se deriva [3.11].

Ejemplo 3.3. Descomposición de la varianza de las puntuaciones empíricas

En la Tabla 3.4 se presentan las medias y las varianzas para las distintas variables.

Tabla 3.4. Medias y varianzas de las puntuaciones *V*, *E* y *X* en las distintas formas

Formas		*Forma* 1		*Forma* 2		*Forma* 3		
	V	E_1	X_1	E_2	X_2	E_3	X_3	
Evaluado 1	12	–2	10	0	12	–1	11	…
Evaluado 2	11	0	11	–2	9	–1	10	…
Evaluado 3	11	0	11	2	13	–2	9	…
Evaluado 4	12	2	14	0	12	0	12	…
Evaluado 5	4	0	4	0	4	0	4	…
…	…	…	…	…	…	…	…	
Medias (μ)	12	0	12	0	12	0	12	
Varianzas (σ^2)	4	2	6	2	6	2	6	

Por ejemplo, μ_V se obtendría como la media de las puntuaciones verdaderas de todos los evaluados de la población (12, 11, 11, 12, 4,…) que es 12. Observe que, en nuestro ejemplo, las medias de todas las formas son iguales entre sí e iguales a la media de las

puntuaciones verdaderas (12). Además, la varianza de las puntuaciones empíricas en cualquiera de las formas (6) es el resultado de sumar a la varianza verdadera (4) la varianza de los errores en esa forma (2). Así pues, la variabilidad de las puntuaciones empíricas en una forma (6) se produce, en parte, por la variabilidad en el verdadero nivel de rasgo (4) y, en parte, por la presencia de errores y su contribución a la variabilidad (2).

Si nuestras formas fueran muy precisas, la varianza de los errores en cada una de ellas sería 0 y la varianza de las puntuaciones empíricas en cada forma sería igual a la varianza de las puntuaciones verdaderas (ver Tabla 3.5). En este caso, el 100 % de la variabilidad de las puntuaciones empíricas en cualquier forma refleja variabilidad en las puntuaciones verdaderas. Las correlaciones entre las puntuaciones verdaderas y empíricas sería 1.

Tabla 3.5. Medias y varianzas de las puntuaciones *V*, *E* y *X* en las distintas formas para un test máximamente preciso

Formas		*Forma* 1		*Forma* 2		*Forma* 3		
	V	E_1	X_1	E_2	X_2	E_3	X_3	
Evaluado 1	6	0	6	0	6	0	6	…
Evaluado 2	11	0	11	0	11	0	11	…
Evaluado 3	11	0	11	0	11	0	11	…
Evaluado 4	12	0	12	0	12	0	12	…
Evaluado 5	4	0	4	0	4	0	4	…
…	…	…	…	…	…	…	…	
Medias (μ)	12	0	12	0	12	0	12	
Varianzas (σ^2)	6	0	6	0	6	0	6	

Puesto que en el Modelo Clásico la varianza de las puntuaciones empíricas se descompone linealmente en varianza verdadera y varianza error sería importante obtener información sobre cuánto de la varianza de las *X* se debe a la varianza de las *V* o saber cuánto correlaciona *X* con *V*. Nos encontramos con el inconveniente de desconocer las auténticas *V* de las *N* personas. En la siguiente sección se ofrece el método propuesto por Spearman para eludir este problema.

Concepto de formas paralelas y coeficiente de fiabilidad

No podemos conocer directamente la correlación entre las puntuaciones empíricas en un test y las puntuaciones verdaderas. Sin embargo, resulta factible obtener la correlación entre las puntuaciones empíricas que proporcionan dos formas paralelas de un test, diseñadas ambas para evaluar el rasgo *V* de las personas. Veremos en esta sección que esta correlación nos proporciona la información que buscamos sobre la fiabilidad de las puntuaciones; es decir, sobre qué proporción de la varianza de *X* se debe a la varianza de *V*.

Hasta ahora hemos estado trabajando con el concepto de formas de un test *X*. El Modelo Clásico original requiere que dichas formas sean *formas paralelas*. Dos formas paralelas X_1 y X_2 de un test se definen como tales mediante dos condiciones:

1. Un individuo tiene la misma puntuación verdadera en ambas formas:

$$V_{i1} = V_{i2} = V_i \qquad [3.12]$$

2. La varianza de los errores de medida en ambas formas es la misma:

$$\sigma^2_{E_1} = \sigma^2_{E_2} \qquad [3.13]$$

Es decir, los dos tests miden con la misma precisión. Las formas que hemos visto en los ejemplos anteriores eran formas paralelas. En lo sucesivo, para simplificar, nos referiremos a la varianza error de cada forma paralela como σ^2_E (donde σ^2_E designa indistintamente a σ^2_{E1} o σ^2_{E2}).

Dos formas suelen hacerse paralelas por diseño, especificando en cada una igual número de ítems y especificaciones similares de contenidos. Por ejemplo, el test formado por los ítems impares de una prueba suele considerarse una forma paralela del test formado por los ítems pares de esa misma prueba (si en principio no hay razón para pensar que los ítems de las dos mitades difieren en conjunto).

Si tres formas (X_1, X_2, y X_3) son paralelas la distribución de sus puntuaciones observadas serán idénticas en cuanto a media (ver [3.10]):

$$\mu_{X_1} = \mu_{X_2} = \mu_{X_3}$$

y varianza (ver [3.11]):

$$\sigma^2_{X_1} = \sigma^2_{X_2} = \sigma^2_{X_3}$$

También serán iguales las covarianzas de esas formas entre sí:

$$\sigma_{X_1X_2} = \sigma_{X_1X_3} = \sigma_{X_2X_3},$$

y las covarianzas con cualquier otra variable *Z*,

$$\sigma_{X_1Z} = \sigma_{X_2Z} = \sigma_{X_3Z}$$

Lo mismo se aplica también a las correlaciones de las formas entre sí:

$$\rho_{X_1X_2} = \rho_{X_1X_3} = \rho_{X_2X_3}$$

Debe observarse que el paralelismo de las formas es lo que las hace intercambiables y lo que dota de significado a la definición operacional de la puntuación verdadera como valor esperado de las puntuaciones a través de las formas.

En lo sucesivo, para simplificar, nos referiremos a cada forma paralela como X (para designar indistintamente a X_1 o X_2), por lo que utilizaremos los términos μ_X y σ^2_X para referirnos a la media y varianza de cualquiera de las formas paralelas.

La correlación entre dos formas paralelas (X_1 y X_2) es muy importante en el Modelo Clásico y se denomina *coeficiente de fiabilidad*. Puesto que ambas formas son paralelas, la correlación nos permite inferir algo sobre el grado de precisión de cualquiera de ellas. Es fácil entender por qué es una medida de precisión. Si las puntuaciones obtenidas en dos formas paralelas son precisas parece razonable esperar una correlación elevada en la población. Si ambas correlacionasen de forma mínima, no podemos fiarnos de que reflejen fidedignamente los niveles de rasgo verdaderos. Por tanto, el coeficiente de fiabilidad es un indicador de precisión; suele utilizarse el símbolo ρ_{XX} y es matemáticamente igual al cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas de cualquiera de las formas:

$$\rho_{XX} \equiv \rho_{X_1X_2} = \frac{\sigma^2_V}{\sigma^2_X} \qquad [3.14]$$

Veamos por qué. La correlación entre formas paralelas puede expresarse como:

$$\rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}$$

Puesto que las formas son paralelas, podemos referirnos con el término σ_X indistintamente a σ_{X_1} *o a* σ_{X_2}:

$$\rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sigma_{X_1X_2}}{\sigma_X\sigma_X} = \frac{\sigma_{X_1X_2}}{\sigma^2_X}$$

Además, $X_1 = V_1 + E_1$ y $X_2 = V_2 + E_2$. Matemáticamente, si tenemos dos variables X_1 y X_2 que son combinación lineal de otras, la covarianza entre ambas se deriva de las covarianzas entre todas las otras:

$$\sigma_{X_1X_2} = \sigma_{V_1V_2} + \sigma_{V_2E_1} + \sigma_{V_1E_2} + \sigma_{E_1E_2}$$

que se puede simplificar, dado [3.7] , [3.8] y [3.12]:

$$\sigma_{X_1X_2} = \sigma^2_V$$

por lo que se llega a la ecuación [3.14]:

$$\rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_X^2} = \frac{\sigma_V^2}{\sigma_X^2}$$

Recordemos que la varianza de las puntuaciones en un test se descomponía en dos componentes, uno relacionado con los errores y otro con las puntuaciones verdaderas (ecuación [3.11]). El valor del coeficiente de fiabilidad puede interpretarse entonces como la proporción de la varianza de las puntuaciones empíricas que puede atribuirse a la variabilidad de las personas en las puntuaciones verdaderas.

Lógicamente, el coeficiente de fiabilidad también nos índica qué proporción de varianza de las puntuaciones en el test no se debe a la varianza de los errores:

$$\rho_{XX} = \frac{\sigma_V^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \qquad [3.15]$$

Nótese además que el coeficiente de fiabilidad puede asumir valores entre 0 y 1, ya que se trata de un cociente de varianzas, cuyo valor es siempre positivo.

Por otro lado, el coeficiente de fiabilidad también se puede interpretar como la correlación al cuadrado entre las puntuaciones verdaderas y las puntuaciones observadas en el test. En efecto:

$$\rho_{XV}^2 = \frac{\sigma_{XV}^2}{\sigma_X^2\sigma_V^2} = \frac{\left(\sigma_V^2 + \sigma_{VE}\right)^2}{\sigma_X^2\sigma_V^2} = \frac{\sigma_V^2}{\sigma_X^2} = \rho_{XX} \qquad [3.16]$$

Es decir, que el coeficiente de fiabilidad es el cuadrado de la correlación entre X y V. Al valor ρ_{XV} se le denomina *índice de fiabilidad*:

$$\rho_{XV} = \sqrt{\rho_{XX}} \qquad [3.17]$$

Tanto el coeficiente como el índice de fiabilidad reflejan la precisión de las medidas siempre que asumamos que en el grupo al que se aplica el test hay cierta variabilidad en la característica que se está midiendo.

Ejemplo 3.4. Interpretación del coeficiente de fiabilidad

Supongamos que la correlación entre dos formas paralelas X_1 y X_2 es 0,67 ($\rho_{X_1X_2} = 0{,}67$) y que la varianza de ambas formas es 6 ($\sigma^2_X = 6$); entonces diríamos que el coeficiente de fiabilidad de las puntuaciones obtenidas en cualquiera de las dos pruebas es 0,67 ($\rho_{XX} = 0{,}67$). Es decir, la correlación entre X_1 y X_2 es el coeficiente de fiabilidad de las puntuaciones en la prueba X_1 (y en la prueba X_2). La varianza de X_1 (o de X_2), en nuestro ejemplo, es 6. El coeficiente de fiabilidad indicaría justamente qué proporción de esos 6 pun-

tos, es varianza verdadera. Puesto que el coeficiente de fiabilidad es 0,67, podemos decir que el 67% de la varianza empírica es varianza verdadera. La varianza verdadera sería, justamente, 4 (el 67% de 6):

$$\sigma_V^2 = \sigma_X^2 \rho_{XX} = 6(0{,}67) = 4$$

También podría deducirse la varianza de los errores de medida, que sería justamente 2 (el 33 % de 6):

$$\sigma_E^2 = \sigma_X^2 (1 - \rho_{XX}) = 6(0{,}33) = 2$$

En nuestro ejemplo, el índice de fiabilidad sería 0,82 (que es la raíz de 0,67).

Fórmula General de Spearman-Brown: Fiabilidad de las puntuaciones en un test formado por *n* formas paralelas

Imaginemos que disponemos de *n* formas paralelas para medir un rasgo psicológico determinado. Según lo visto, las *n* formas tendrán en la población las mismas varianzas empíricas. Además, las correlaciones entre todos los posibles pares de formas paralelas que podemos establecer serán también iguales, e indicarán la fiabilidad de cualquiera de ellas a la hora de determinar los niveles de rasgo. Estudiemos las propiedades psicométricas de un test que es el resultado de unir varias formas paralelas.

Denominemos las puntuaciones originales de cada persona en las *n* formas paralelas como $X_1,\ldots, X_n$, $V_1,\ldots, V_n$, $E_1,\ldots$ y E_n. Las puntuaciones en el test final alargado se obtienen sumando las puntuaciones en las *n* formas:

$$X_a = X_1 + \ldots + X_n \qquad [3.18]$$

Para cada forma paralela, podemos separar la parte verdadera y la parte error: $X_a = V + E_1 + \ldots + V + E_n$; así podemos definir $X_a = V_a + E_a$, donde $V_a = nV$ y $E_a = E_1 + \ldots + E_n$, ya que la puntuación verdadera es la misma en cada forma paralela, mientras que el error puede cambiar de una forma a otra. Los parámetros de la población en una forma paralela (cualquiera de ellas) podemos designarlos como σ^2_X, σ^2_V, σ^2_E y ρ_{XX}. Si unimos *n* formas paralelas en un único test, los parámetros de este test alargado podemos expresarlos como σ^2_{Xa}, σ^2_{Va}, σ^2_{Ea} y ρ_{nXX}. Vamos a llegar a determinadas expresiones para obtener los parámetros del test alargado conociendo los parámetros de una forma paralela.

La varianza empírica del test formado por *n* formas paralelas será:

$$\sigma^2_{Xa} = n\sigma^2_X + n(n-1)\sigma^2_X \rho_{XX} = n\sigma^2_X \left[1 + (n-1)\rho_{XX}\right] \qquad [3.19]$$

La varianza verdadera del test formado por *n* formas paralelas será:

$$\sigma^2_{Va} = n^2\sigma^2_V \qquad [3.20]$$

puesto que V_a es una transformación lineal de V ($V_a = nV$, donde n es una constante).
La varianza error del test formado por *n* formas paralelas será:

$$\sigma^2_{Ea} = n\sigma^2_E + n(n-1)\sigma^2_E\rho_{EE} = n\sigma^2_E \qquad [3.21]$$

ya que ρ_{EE} , la correlación entre los errores de dos formas, es 0 según el 4º supuesto.

A partir de las expresiones anteriores, y recordando que el coeficiente de fiabilidad es el cociente entre la varianza verdadera y la varianza empírica, podemos obtener el coeficiente de fiabilidad de las puntuaciones en un test alargado *n* veces (ρ_{nXX}):

$$\rho_{nxx} = \frac{\sigma^2_{Va}}{\sigma^2_{Xa}} = \frac{n^2\sigma^2_V}{n\sigma^2_{XX}\left[1+(n-1)\rho_{XX}\right]} = \frac{n\rho_{XX}}{1+(n-1)\rho_{XX}} \qquad [3.22]$$

La expresión [3.22] se conoce como *Fórmula General de Spearman-Brown*, y permite obtener el coeficiente de fiabilidad de las puntuaciones en un test compuesto por *n* formas paralelas (es decir, cuál será el coeficiente de fiabilidad, ρ_{nXX}, de un test que se forma con *n* versiones paralelas de un test inicial que tiene un coeficiente de fiabilidad, ρ_{XX}).

Ejemplo 3.5. Fiabilidad de las puntuaciones en un test formado por n formas paralelas

Si formamos un nuevo test uniendo las dos formas paralelas X_1 y X_2, ambas con varianzas iguales ($\sigma^2_X = 6$, $\sigma^2_V = 4$, $\sigma^2_E = 2$) e igual coeficiente de fiabilidad ($\rho_{XX} = 0{,}67$), se obtiene un nuevo test con varianzas:

$$\sigma^2_{Xa} = n\sigma^2_X\left[1+(n-1)\rho_{XX}\right] = (2)6\left[1+(1)0{,}67\right] = 20$$
$$\sigma^2_{Va} = n^2\sigma^2_V = 2^2(4) = 16$$
$$\sigma^2_{Ea} = n\sigma^2_E = 2(2) = 4$$

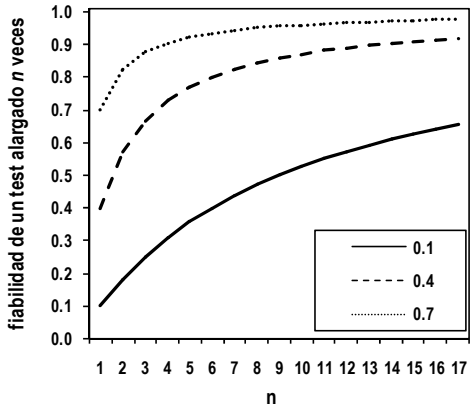
El coeficiente de fiabilidad del nuevo test sería:

$$\rho_{nXX} = \frac{n\rho_{XX}}{1+(n-1)\rho_{XX}} = \frac{(2)0{,}67}{1+(1)0{,}67} = 0{,}8$$

El 80% de la varianza del nuevo test ($\sigma^2_{Xa} = 20$) es varianza verdadera ($\sigma^2_{Va} = 16$).

Según la Fórmula General de Spearman-Brown, el coeficiente de fiabilidad aumenta al alargar un test. Esto ocurre porque, al añadir $n - 1$ formas paralelas, la varianza debida a las puntuaciones verdaderas se incrementa más rápido ($\sigma^2_{Va} = n^2\sigma^2_V$) que la varianza debida a los errores ($\sigma^2_{Ea} = n\,\sigma^2_E$). La Figura 3.1 muestra el efecto de multiplicar la longitud del test por n (n: 1, 2, ... 17) en 3 tests que difieren originalmente en ρ_{XX} (0,1, 0,4 y 0,7).

Figura 3.1. Coeficiente de fiabilidad de las puntuaciones en un test alargado como función del coeficiente de fiabilidad del test original y del número *n* de formas paralelas



El valor $n = 1$ representa lo que ocurre con el test original. Los otros valores de n (2, 3,…) se refieren a lo que ocurre con los tests alargados (de longitud duplicada, triplicada,…). Observe, que en cualquiera de los tres tests, el incremento de la fiabilidad a medida que se incrementa la longitud no es lineal. Esto quiere decir, por ejemplo, que al pasar de $n = 1$ a $n = 2$ se obtiene mayor ganancia en precisión que al pasar de $n = 2$ a $n = 3$. También puede observarse que el incremento posible será menor cuanto mayor sea el coeficiente de fiabilidad del test original, ya que el valor máximo que puede obtenerse es 1.

Ejemplo 3.6. Fiabilidad de las puntuaciones en un test formado por n formas paralelas

Supongamos que una prueba de atención de 25 ítems obtiene en un grupo normativo un $\rho_{XX} = 0,7$. Si se añadieran 75 ítems (tres formas paralelas) al test inicial, el test alargado tendría 100 ítems (4 veces el inicial), y su fiabilidad sería:

$$\rho_{nXX} = \frac{n\rho_{XX}}{1 + (n-1)\rho_{XX}} = \frac{(4)0,7}{1 + (3)0,7} = 0,903$$

Si cuadriplicáramos la longitud del test recién formado, pasaríamos de 100 a 400 ítems. El test alargado tendría un coeficiente de fiabilidad:

$$\rho_{nXX} = \frac{n\rho_{XX}}{1+(n-1)\rho_{XX}} = \frac{(4)0{,}903}{1+(3)0{,}903} = 0{,}974$$

En el primer caso, el incremento que se produce al multiplicar por 4 la longitud inicial del test de atención es de 0,203, mientras que en el segundo caso, el incremento es únicamente de 0,071 (a pesar de que en el primer caso hemos añadido 75 ítems y en el segundo 300!). Esto se debe a que el coeficiente de fiabilidad del test inicial es mayor en el segundo caso que en el primero y a que el segundo test tiene ya un considerable número de ítems (100).

Es importante tener en cuenta que la Formula General de Spearman-Brown no debe aplicarse cuando las formas añadidas no son paralelas o cuando al incrementar la longitud se producen efectos de fatiga (o de la práctica) al responder. Esto último ocurre, por ejemplo, cuando la persona no responde a los nuevos ítems con igual motivación, eficacia, atención, etc. Tampoco conviene olvidar que se requiere que la prueba original haya sido aplicada a un número suficiente de sujetos, de forma que el coeficiente de fiabilidad se halle bien estimado (Alsawalmeh y Feldt, 1999). Establecidas estas limitaciones, la fórmula de Spearman-Brown puede utilizarse para:

1. Extrapolar cual sería el número de ítems necesarios para que las puntuaciones en nuestro instrumento alcancen una determinada fiabilidad. Así, despejando n de la fórmula anterior:

$$n = \frac{\rho_{nXX}(1-\rho_{XX})}{\rho_{XX}(1-\rho_{nXX})} \qquad [3.23]$$

   donde ρ_{nXX} indica la fiabilidad que se quiere obtener, ρ_{XX} indica la fiabilidad actual y n es el número de formas paralelas que debería tener el test final para que se alcance esa fiabilidad. Lógicamente, si el test original tiene J ítems el test final deberá tener J' ítems, donde J' es igual a:

$$J' = nJ$$

   En la práctica, la ecuación [3.23] puede resultar eficaz para diseñar un test inicial corto y estimar cuál debería ser su longitud para obtener un coeficiente de fiabilidad determinado, y así comprobar si merece la pena continuar con el diseño de nuevos ítems paralelos o reformar los ya creados.
2. Para poder comparar la fiabilidad de las puntuaciones en dos pruebas con distinto número de ítems. Si una prueba tiene J ítems y otra tiene J' podemos ver cuál sería la fiabilidad de la primera si tuviera J' ítems ($J' > J$), para ello basta utilizar la fórmula de Spearman-Brown sustituyendo n por J'/J.

En ambas situaciones, n siempre indica el número de veces que el test final contiene al test original y $n - 1$ indica el número de formas que se añaden a la forma original.

Ejemplo 3.7. Utilidad de la fórmula de Spearman-Brown

Supongamos que para las puntuaciones en un test inicial de 25 ítems se obtiene un coeficiente de fiabilidad de 0,65, considerado bajo para los objetivos que se pretenden conseguir con su aplicación. Una manera de incrementar su precisión es alargarlo con ítems paralelos a los iniciales. Al constructor de la prueba le interesa que el test tenga, al menos, un coeficiente de fiabilidad de 0,86, y se pregunta con cuántos ítems lo conseguiría. Aplicando la fórmula [3.23], obtenemos:

$$n = \frac{\rho_{nxx}(1-\rho_{xx})}{\rho_{xx}(1-\rho_{nxx})} = \frac{0{,}86(1-0{,}65)}{0{,}65(1-0{,}86)} = 3{,}308$$

Esto significa que si multiplicamos por 3,308 la longitud inicial del test, es decir, con un test de 83 ítems (3,308(25) = 82,7), conseguiremos la precisión deseada. Por tanto, a los 25 ítems que tiene el test inicial habría que añadir 58 ítems paralelos (2,308 formas) para conseguir la fiabilidad de 0,86.

Otro ejemplo. Consideremos que dos pruebas tienen, respectivamente, coeficientes de fiabilidad 0,65 y 0,7. La primera tiene 15 ítems y la segunda 20. ¿Cuál de las dos pruebas sería más precisa si ambas tuvieran el mismo número de ítems? Para responder a esta pregunta, podemos calcular cúal sería el coeficiente de fiabilidad de las puntuaciones en la primera prueba si tuviera 20 ítems:

$$n = \frac{J'}{J} = \frac{20}{15} = 1{,}33$$

El coeficiente de fiabilidad de la primera prueba sería:

$$\rho_{nxx} = \frac{n\rho_{xx}}{1+(n-1)\rho_{xx}} = \frac{1{,}33(0{,}65)}{1+0{,}33(0{,}65)} = 0{,}712$$

Lo que significa que, con el mismo número de ítems, la primera prueba sería más fiable en la muestra.

Aproximaciones a la fiabilidad y tipos de error

Hemos visto que, a partir del Modelo Clásico, se expresa un nuevo concepto, la fiabilidad de las puntuaciones en el test, que representa la proporción de la varianza de las puntua-

ciones en el test que se debe a la varianza de las puntuaciones verdaderas. En términos generales puede considerarse que la fiabilidad nos indica la replicabilidad de la medida a través de distintas condiciones, momentos, formas del test, etc. Ahora bien, la visión que se ha dado hasta ahora (coeficiente de fiabilidad como correlación entre formas paralelas) se encuentra algo simplificada. En realidad, el concepto de fiabilidad (o replicabilidad) de las puntuaciones es más complejo y puede entenderse de distintas maneras:

1. Ya hemos observado que podemos calcular el coeficiente de fiabilidad como una correlación entre formas paralelas. En ese caso estaríamos estudiando si se replican las mismas medidas al aplicar una prueba paralela con ítems distintos. Si ambas formas son paralelas, la correlación entre ambas indica su grado de *equivalencia*. En este sentido, replicabilidad implica que debemos obtener las mismas medidas cuando medimos lo mismo con pruebas equivalentes.
2. También puede aludirse a la *estabilidad* temporal de las medidas que proporciona nuestro instrumento. En este sentido, replicabilidad implica que debemos obtener las mismas medidas cuando medimos lo mismo en momentos distintos.
3. Finalmente, puede hacerse referencia al grado en que diferentes partes del test miden un rasgo con *consistencia*. En este sentido, replicabilidad implica que debemos obtener las mismas puntuaciones cuando medimos lo mismo con distintas partes del test.

Según el procedimiento utilizado para calcular el coeficiente de fiabilidad estaremos siendo sensibles en mayor o menor grado a distintas fuentes de error. Ya hemos observado que, en el Modelo Clásico, se establece que:

$$X_f = V + E_f$$

Y también que el error es el resultado de todos aquellos factores (de la persona, de la situación o relativos a la composición del test) que hacen que la puntuación observada de una persona se aleje de su valor esperado. De forma más sencilla, un error implica un cambio en la puntuación de una persona de una medición a otra. Algunos autores suelen distinguir entre tres tipos de fuentes de error en los tests de respuesta seleccionada (Schmidt y Hunter, 1996, 1999; Schmidt, Le y Ilies, 2003):

1. Los *errores debidos a factores transitorios* suponen cambios en las respuestas de una persona que se deben a factores que cambian de una sesión de aplicación a otra pero que, dentro de una sesión, afectan por igual a todos los ítems. Si aplicamos un test dos veces puede haber cambios en variables personales (salud, humor, motivación, eficiencia mental, concentración, minuciosidad, impulsividad, etc.) o en variables situacionales que no han sido controladas en la aplicación (claridad de las instrucciones, presencia de incentivos, tiempo de la aplicación, etc.). Tales variables pueden tener efectos en todas las medidas tomadas dentro de una misma sesión de aplicación. Cambios en esas variables a través de las sesiones producirán cambios en las puntuaciones observadas. Por ejemplo, si alguien responde a una prueba de conocimientos con baja motivación, su rendimiento puede verse afectado. Su puntuación X estará por debajo de su puntuación V. Si volvemos a aplicar el mismo test en otro momento, en el que tenga mayor

motivación, su puntuación *X* será mayor. El nivel de motivación afectará a todas sus respuestas recogidas en cada sesión.

Los errores debidos a factores transitorios pueden detectarse estudiando cómo varía el rendimiento de la persona en distintos momentos temporales. Estos errores no son detectables si el test se aplica una sola vez (pues en ese caso no podemos saber cómo cambiarían las puntuaciones de las personas si se les aplica el test en otro momento).

2. Los *errores debidos a la especificidad* suponen cambios en las respuestas de una persona que se deben al contenido concreto de los ítems que se le presentan. Por ejemplo, en una prueba de conocimientos de Filosofía cada estudiante puede tener distinto nivel de dominio de los distintos temas. Alguien puede saber mucho de Platón y poco de Kant. Si le hiciéramos una pregunta sobre Platón su puntuación *V* se sobrestimaría (y ocurriría lo inverso si le hiciéramos una pregunta sobre Kant). Otro ejemplo: en una escala de Estabilidad Emocional se incluyen ítems en sentido directo y otros en sentido inverso; las respuestas de una persona en ítems de uno y otro tipo pueden ser distintas.

   Los errores debidos a la especificidad pueden detectarse estudiando cómo varía el rendimiento de la persona en distintas partes del test. Estos errores no son detectables si se aplica la misma pregunta en dos ocasiones distintas (pues en ese caso no podemos saber cómo cambiarían las puntuaciones de las personas si les hubiéramos hecho otra pregunta).

3. Los *errores debidos a factores aleatorios* se refieren al grado de inconsistencia en la respuesta que no puede ser atribuido directamente al contenido de los ítems, ni a otros factores de la persona o la situación que actúan de forma sistemática en el tiempo. Es producto de variaciones en la atención, de distracciones momentáneas, de la propia labilidad intrínseca a nuestro sistema nervioso, etc. Por ejemplo, supongamos que al escuchar un ítem de una prueba de inglés, el evaluado se ha distraído; como no ha escuchado la pregunta, falla el ítem. Consideremos que de haber escuchado el ítem lo hubiera acertado. En ese caso, si se le vuelve a aplicar el ítem lo acertará. La distracción no produce un error debido a la especificidad del ítem ya que el fallo inicial de la persona no tiene que ver con el contenido del ítem. Tampoco es un error debido a factores transitorios ya que la distracción no necesariamente ha afectado a todos los ítems aplicados en la misma sesión.

En los siguientes apartados se recoge el grado en que cada coeficiente de fiabilidad es sensible a cada tipo de error. Antes de empezar, es necesario advertir de un cambio de notación. Hasta el momento, el Modelo Clásico y los estadísticos (medias, varianzas, correlaciones,...) se han descrito en términos paramétricos; es decir, para la población. Por ello, se utilizaba la nomenclatura griega (σ^2_X , ρ_{xx} , etc.). En la práctica vamos a disponer de datos obtenidos en una muestra o grupo normativo concreto (S^2_X , r_{xx}, etc.). Esto significa que, de modo directo, únicamente vamos a disponer de las puntuaciones empíricas de dicha muestra, a partir de las cuales podemos obtener los estadísticos que sean oportunos.

Fiabilidad como correlación entre formas paralelas

A veces, por razones de índole práctica o investigadora, se diseña un test y una segunda versión del mismo, denominada forma paralela, que intenta evaluar o medir lo mismo que el test original pero con diferentes ítems. Como ya hemos explicado, dos versiones o formas se consideran paralelas si su contenido es similar y, aplicadas a una misma muestra de personas, obtienen similares medias, varianzas y covarianzas con otras variables.

La correlación de Pearson entre las puntuaciones obtenidas en una misma muestra en dos formas paralelas se considera el coeficiente de fiabilidad de cualquiera de ellas, e indicará el grado en que pueden considerarse equivalentes. Por ello, en ocasiones se denomina a este coeficiente de fiabilidad *coeficiente de equivalencia*. Si las formas no fuesen paralelas puede subestimarse dicho coeficiente.

Ejemplo 3.8. Coeficiente de fiabilidad por el método de las formas paralelas

Se han aplicado las dos formas de un test a 13 personas[3], obteniendo los resultados que se muestran en la Tabla 3.6.

Tabla 3.6. Cálculo del coeficiente de fiabilidad por el método de las formas paralelas

Evaluados	1	2	3	4	5	6	7	8	9	10	11	12	13
X_1 = *Forma* 1	10	12	11	14	11	9	13	14	16	15	13	14	16
X_2 = *Forma* 2	9	13	14	16	10	10	13	15	14	14	12	10	17

El coeficiente de fiabilidad sería:

$$r_{XX} = r_{X_1X_2} = 0{,}701$$

Lo que se significa que el 70,1 % de la varianza de las puntuaciones observadas en cualquiera de las formas es debida a la variabilidad en el verdadero nivel de rasgo.

Es evidente que la falta de concordancia cuando se calcula el coeficiente de equivalencia se deberá a que las dos formas tienen contenidos específicos distintos o a que los ítems no son adecuados para medir el rasgo. En sentido estricto, el coeficiente de equivalencia no es sensible a los errores debidos a factores transitorios y nos indica qué proporción de la varianza de las puntuaciones empíricas en el test completo no se debe a la varianza de los errores debidos a factores específicos o al error debido a factores aleatorios. Sin embargo,

[3] El uso de sólo 13 sujetos para evaluar las propiedades psicométricas de una prueba es, lógicamente, insuficiente. El pequeño tamaño de la muestra se debe a razones puramente didácticas, de forma que el lector pueda realizar los cálculos si lo desea.

su interpretación suele simplificarse, de tal forma que se entiende como proporción de varianza de las puntuaciones empíricas explicada por las puntuaciones verdaderas. Esta última interpretación es razonable para ciertos rasgos, como por ejemplo el nivel de vocabulario, para los que se espera un efecto pequeño de los factores transitorios (ver por ejemplo, Reeve, Heggestad y George, 2005).

No es común diseñar una forma paralela de un test para obtener datos sobre su fiabilidad. Cuando se diseñan (tarea por otra parte difícil) es porque van a utilizarse en determinados trabajos que requieren dos aplicaciones sucesivas de un test cuyos contenidos se pueden recordar con facilidad. Por ejemplo, para evaluar la eficacia de ciertos programas cortos de enriquecimiento cognitivo o motivacional, conviene utilizar antes y después del entrenamiento pruebas equivalentes aunque con contenidos diferentes (formas paralelas), para evitar los efectos del recuerdo.

Fiabilidad como estabilidad temporal

Si disponemos de las puntuaciones de N personas en un test y, después de transcurrido un tiempo, volvemos a medir a las mismas personas en el mismo test, cabe esperar una correlación de Pearson elevada entre ambas mediciones (reflejando así la concordancia de las medidas tomadas en dos momentos distintos). Dicha correlación entre la evaluación test y la evaluación retest ($r_{XtestXretest}$) se denomina *coeficiente de fiabilidad test-retest o de estabilidad temporal*, e indicará tanta mayor estabilidad temporal de las puntuaciones en la prueba cuanto más cercano a uno sea.

Este modo de operar se desprende también directamente del Modelo Clásico, según el cuál se define la fiabilidad como la correlación entre las puntuaciones empíricas en dos formas paralelas, ya que no existe mayor grado de paralelismo entre dos tests que cuando en realidad es uno aplicado dos veces.

Ejemplo 3.9. Coeficiente de fiabilidad por el método test-retest

Para obtener el coeficiente de estabilidad de una escala se aplica una forma del test a una muestra. Transcurridos dos meses, se vuelve a aplicar la misma forma a las mismas personas bajo las mismas condiciones. Sus puntuaciones directas en las dos aplicaciones son las que aparecen en la Tabla 3.7.

Tabla 3.7. Cálculo del coeficiente de estabilidad

Evaluados	1	2	3	4	5	6	7	8	9	10	11	12	13
X_{Test}	10	12	11	14	11	9	13	14	16	15	13	14	16
X_{Restest}	11	12	13	15	12	12	10	15	13	18	11	15	17

Para obtener el coeficiente de fiabilidad test-retest bastaría con correlacionar los datos de las dos aplicaciones:

$$r_{X_{test}X_{tetest}} = 0{,}639$$

En este caso se obtiene una cierta estabilidad de las puntuaciones. Si los niveles de rasgo de las personas no han variado a lo largo de los dos meses transcurridos entre las dos aplicaciones, podemos decir que el test proporciona ciertas garantías (no óptimas) respecto a la precisión con la que mide, dado que una persona concreta obtiene puntuaciones muy parecidas (o similares) en las dos aplicaciones.

Este coeficiente se obtiene, sobre todo, en pruebas cuyo objetivo de medida es un rasgo estable (pruebas de inteligencia general, aptitudes, rasgos de personalidad, etc.) dado que, de lo contrario, no se podría discernir entre la inestabilidad debida al rasgo de la causada por la falta de precisión del instrumento. Es decir, es necesario asumir que las puntuaciones verdaderas de los evaluados no han cambiado entre el test y el retest. Por tanto, no es adecuado calcular este coeficiente para cuando se pretenden medir atributos psicológicos que por naturaleza son fluctuantes (p.ej., estados de ansiedad).

La determinación del intervalo temporal entre aplicaciones es importante y debe ser informada (Standards, AERA, APA y NCME, 1999; p. 32). Para establecer un período concreto, el efecto en las respuestas debido a la doble aplicación (efectos del aprendizaje, la fatiga, la maduración, el recuerdo, la motivación, el deseo de congruencia, etc.) debería ser analizado y controlado. Un efecto debido a la doble aplicación implicaría que: (1) las puntuaciones verdaderas de las personas han cambiado; (2) la precisión de las medidas ha variado entre el test y el retest. Si el intervalo es demasiado corto y no hay efectos de fatiga suele producirse una sobreestimación de la fiabilidad porque se recuerdan las respuestas. Por tanto, es aconsejable dejar periodos largos cuando los ítems y las respuestas pueden memorizarse con facilidad; de lo contrario, los evaluados podrían emitir pautas de respuesta similares en las dos aplicaciones del test únicamente por efectos del recuerdo y del deseo de responder de manera congruente. Debe tenerse en cuenta, sin embargo, que cuanto mayor es el intervalo temporal que se deja entre ambas aplicaciones, mayor es la posibilidad de que ocurran cambios reales en el rasgo (p.ej., por factores de tipo madurativo) y, por lo tanto, se subestime la fiabilidad de la prueba. El intervalo usual suele variar entre dos semanas y dos meses.

Por otro lado, es importante tener en cuenta que, dado que se aplica la misma forma (i.e., las mismas preguntas) en dos momentos distintos, este coeficiente de fiabilidad no es sensible a los errores debidos a la especificidad. En sentido estricto, el coeficiente de estabilidad nos indica qué proporción de la varianza de las puntuaciones empíricas en el test completo no se debe a la varianza de los errores debidos a factores transitorios o al error debido a factores aleatorios. Para la mayoría de los rasgos (p.ej., neuroticismo, capacidad verbal, etc.) el efecto de los errores debidos a la especificidad es importante. Por tanto, el coeficiente de estabilidad no puede ser considerado una buena estimación de la proporción de varianza de las puntuaciones empíricas que es explicada por las puntuaciones verdaderas.

Fiabilidad como consistencia interna

También se han propuesto otros coeficientes basados en una única aplicación del test y que, por tanto, son menos costosos de obtener. Con estos métodos se estudia la concordancia entre las puntuaciones de los evaluados en distintas partes del test. Así, la fiabilidad se entiende ahora como el grado en que diferentes subconjuntos de items covarían, correlacionan o son consistentes entre sí. Todos estos coeficientes no son, por tanto, sensibles al los errores debidos a factores transitorios.

Lo más usual es estudiar la consistencia entre las dos mitades del test (método de dos mitades) o entre tantas partes como elementos tenga la prueba (consistencia interna global).

Coeficiente de fiabilidad por el método de las dos mitades (método de Spearman-Brown)

En primer lugar se divide el test en dos mitades (p.ej., ítems impares e ítems pares). Para cada persona se obtiene la puntuación directa en ambas mitades. Disponemos entonces de dos variables (X_I y X_P), cuya correlación de Pearson indica su grado de relación lineal. Si ambas mitades son paralelas, su correlación será el coeficiente de fiabilidad de las puntuaciones en la mitad del test. Una práctica habitual consiste en extrapolar el coeficiente de fiabilidad de las puntuaciones en el test completo ($X = X_I + X_P$) aplicando la fórmula de Spearman-Brown (haciendo $n = 2$ ya que el test completo tiene el doble de items que cualquiera de sus mitades):

$$_{SB}r_{XX} = \frac{2r_{X_I X_P}}{1 + r_{X_I X_P}} \qquad [3.24]$$

A partir de esta fórmula podemos comprobar que el coeficiente de fiabilidad, entendido como la expresión de la consistencia entre dos mitades, es mayor que la correlación de Pearson entre ambas mitades. Sus valores pueden estar entre 0 y 1 e indica el grado en que un test formado por dos formas paralelas (las mitades) proporcionaría resultados similares a otro test equivalente. En sentido estricto, nos indica qué proporción de la varianza de las puntuaciones empíricas en el test completo no se debe a la varianza error por muestreo de contenidos o error aleatorio. Sin embargo, suele interpretarse como proporción de varianza de las puntuaciones en el test que es debida a las puntuaciones verdaderas.

Ejemplo 3.10. Coeficiente de fiabilidad por el método de las dos mitades

En la Tabla 3.8 se muestran los resultados de una muestra de 10 evaluados que responden a un test de 6 ítems (X_1, X_2, …, X_6) valorados de forma dicotómica. En este caso se obtendría que $r_{X_I X_P} = 0{,}277$, y por tanto:

$$_{SB}r_{XX} = \frac{2(0{,}277)}{1+0{,}277} = 0{,}434$$

De nuevo el tope de r_{XX} lo tenemos en 1, con lo que podemos decir que las dos mitades del test no son muy consistentes entre sí. Únicamente un 43.3 % de la varianza de las puntuaciones empíricas se debe a la varianza de las verdaderas. No podríamos afirmar con suficiente certeza que ambas mitades miden con precisión el rasgo de interés.

Tabla 3.8. Cálculo del coeficiente de fiabilidad por el método de las dos mitades

Evaluados	X_1	X_2	X_3	X_4	X_5	X_6	X_I	X_P	X
1	1	0	1	0	1	0	3	0	3
2	0	1	1	1	0	1	1	3	4
3	0	0	1	0	0	0	1	0	1
4	0	1	1	1	0	0	1	2	3
5	0	0	0	1	0	0	0	1	1
6	1	1	1	1	1	1	3	3	6
7	1	1	1	1	1	1	3	3	6
8	0	1	1	1	0	1	1	3	4
9	0	1	0	0	0	0	0	1	1
10	0	0	0	1	0	0	0	1	1
Varianza (S^2)	0,233	0,267	0,233	0,233	0,233	0,267	1,567	1,567	4

Al calcular el coeficiente de fiabilidad por el método de las dos mitades hay que tener en cuenta varias precauciones:

1. La razón de dividir el test en la mitad par y la impar es garantizar su equivalencia. Los tests de rendimiento óptimo suelen tener ítems ordenados en dificultad, de tal forma que se comienza a responder los ítems más fáciles hasta llegar a los situados al final del test, que son los más difíciles. Si realizásemos la partición en dos mitades atendiendo a su disposición en la prueba (la primera mitad formada por los primeros *J*/2 ítems, la segunda por los *J*/2 ítems últimos) difícilmente podría cumplirse que ambas tuvieran la misma media. Por ello, para obtener este coeficiente, hay que cuidar el modo en que se forman las mitades para garantizar su paralelismo; así, ambas mitades deberían estar equilibradas en cuanto a la dificultad, los contenidos y la posición media de sus ítems en el test. Por ejemplo, si en un test de 20 ítems de Extraversión, 10 miden Sociabilidad

y otros 10 miden Impulsividad, las dos formas que construyamos deberían estar equilibradas en el número de ítems de ambas facetas.

2. Si las respuestas a los ítems dependen demasiado de su orden serial en el test (p.ej., en un test de velocidad) es preferible utilizar otros procedimientos para evitar que el coeficiente de fiabilidad se sobrestime. También hay que ser cautos cuando existen grupos de ítems que hacen referencia a un estímulo común (testlets); al repartir esos ítems a través de las mitades, se puede sobrestimar el coeficiente de fiabilidad. En esos casos, el hecho de que una persona obtenga la misma puntuación en las dos partes podrá ser considerado un artefacto metodológico (p.ej., en una prueba de velocidad, la puntuación en la parte del test formada por los ítems impares siempre será muy similar a la puntuación en la parte del test formada por los ítems pares). En el caso de pruebas de velocidad se recomienda no utilizar índices de consistencia interna (o proceder a la eliminación del análisis de los ítems que no han sido alcanzados por un porcentaje de personas). En el caso de ítems que hacen referencia a un estímulo común se recomienda que se mantengan en una misma mitad, ya que si se reparten entre las dos mitades se sobrestimará el coeficiente de fiabilidad (ver Haertel, 2006).
3. Un inconveniente de este método es que existen muchas formas de dividir el test en dos mitades y cada una de ellas arrojará un resultado distinto. De hecho, para McDonald (1999) el procedimiento de las dos mitades no es recomendable porque introduce en su estimación la variabilidad debida al método utilizado para dividir el test en dos mitades.

Al calcular el coeficiente de fiabilidad mediante la fórmula de Spearman-Brown se asume que las dos mitades son formas paralelas. Esto no ocurrirá cuando las dos mitades difieran en el número de ítems (variarán las varianzas verdaderas y las varianzas de error). Por ejemplo, cuando el número de ítems es impar, es incorrecto aplicar la fórmula de Spearman-Brown directamente, puesto que las dos fórmas ya no serían paralelas. En ese caso, el coeficiente de fiabilidad obtenido por la fórmula de Spearman-Brown supone una pequeña subestimación del coeficiente de fiabilidad, por lo que puede calcularse un coeficiente de fiabilidad corregido. En el capítulo 10 se muestran otras formas de calcular el coeficiente por el método de las dos mitades cuando las formas no son paralelas.

Coeficiente α de Cronbach

Como hemos indicado, existen muchas formas de dividir el test en dos mitades. Para resolver este problema se ha propuesto el *coeficiente alfa*, un indicador de consistencia interna con el que se estudia la concordancia entre las puntuaciones de las personas entre las partes más elementales del test: los ítems.

Considere que tenemos un test formado por J ítems:

$$X = \sum_j X_j = \sum_j V_j + \sum_j E_j \qquad [3.25]$$

Si se cumplen los supuestos del Modelo Clásico podemos definir la proporción de varianza del test que es varianza verdadera:

$$\frac{\sigma_V^2}{\sigma_X^2} = \frac{\sum_j \sigma_{V_j}^2 + \sum_{j \neq j'} \sigma_{V_j V_{j'}}}{\sigma_X^2} \qquad [3.26]$$

Los parámetros que aparecen en el numerador se refieren a las puntuaciones verdaderas en los ítems. Sin embargo, asumiendo los supuestos de la TCT, se cumplirá poblacionalmente que el promedio de las covarianzas empíricas entre ítems es igual al promedio de las covarianzas verdaderas:

$$\frac{\sum_{j \neq j'} \sigma_{X_j X_{j'}}}{J(J-1)} = \frac{\sum_{j \neq j'} \sigma_{V_j V_{j'}}}{J(J-1)} \qquad [3.27]$$

y, por tanto:

$$\sum_{j \neq j'} \sigma_{V_j V_{j'}} = \sum_{j \neq j'} \sigma_{X_j X_{j'}} \qquad [3.28]$$

Por otro lado, la covarianza entre dos variables nunca puede ser mayor que la varianza de cualquiera de ellas; por tanto, debe cumplirse siempre que el promedio de las covarianzas verdaderas entre ítems es menor o igual que el promedio de sus varianzas verdaderas:

$$\frac{\sum_{j \neq j'} \sigma_{V_j V_{j'}}}{J(J-1)} \leq \frac{\sum_j \sigma_{V_j}^2}{J} \qquad [3.29]$$

Y, por tanto, considerando [3.28] y [3.29]:

$$\sum_j \sigma_{V_j}^2 \geq \frac{\sum_{j \neq j'} \sigma_{V_j V_{j'}}}{J-1} = \frac{\sum_{j \neq j'} \sigma_{X_j X_{j'}}}{J-1} \qquad [3.30]$$

De las ecuaciones [3.26], [3.28] y [3.30] se deriva la siguiente relación:

$$\frac{\sigma_V^2}{\sigma_X^2} = \frac{\sum_j \sigma_{V_j}^2 + \sum_{j \neq j'} \sigma_{V_j V_{j'}}}{\sigma_X^2} \geq \frac{\frac{\sum_{j \neq j'} \sigma_{X_j X_{j'}}}{(J-1)} + \sum_{j \neq j'} \sigma_{X_j X_{j'}}}{\sigma_X^2} = \frac{J}{J-1} \frac{\sum_{j \neq j'} \sigma_{X_j X_{j'}}}{\sigma_X^2} \qquad [3.31]$$

Pues bien, el denominado coeficiente α (Cronbach, 1951) es:

$$\alpha = \frac{J}{J-1}\frac{\sum_{j\neq j'}\sigma_{X_jX_{j'}}}{\sigma_X^2} \qquad [3.32]$$

Y se dice que el coeficiente α es un límite inferior del coeficiente de fiabilidad ya que siempre toma valores iguales o por debajo de este [$\alpha \leq \sigma_V^2 / \sigma_X^2$].
Para datos muestrales tres formas de expresar el coeficiente α serían[4]:

Ecuación 1 [3.33]

$$\alpha = \frac{J}{J-1}\left(\frac{\sum_{j\neq j'}S_{X_jX_l}}{S_X^2}\right)$$

Ecuación 2 [3.34]

$$\alpha = \frac{J}{J-1}\left(1-\frac{\sum_{j}S_{X_j}^2}{S_X^2}\right)$$

Ecuación 3 [3.35]

$$\alpha = \frac{\overline{S}_{X_jX_{j'}}}{\overline{S}}$$

donde $\sum_{j\neq j'}S_{X_jX_l}$ es la suma de las covarianzas entre ítems, $\sum_{j}S_{X_j}^2$ es la suma de las varianzas de los ítems, $\overline{S}_{X_jX_{j'}}$ indica el promedio de las covarianzas entre ítems $\overline{S}_{X_jX_l} = \left(\sum_{j\neq j'}S_{X_jX_l}\right) \Big/ \left(J(J-1)\right)$ y $\overline{S}$ es el promedio de los J^2 elementos de la matriz de varianzas-covarianzas entre ítems: $\overline{S} = \left(\sum_{j}S_{X_j}^2 + \sum_{j\neq j'}S_{X_jX_l}\right) \Big/ J^2$.

El coeficiente α es útil para expresar en qué grado las medidas que obtenemos de las personas dependen de los ítems aplicados. La pregunta a la que se responde es: ¿los evaluados habrían obtenido puntuaciones similares si hubiéramos aplicado otro test de la misma longitud construido siguiendo la misma lógica?

El coeficiente α siempre toma valores menores o iguales a 1 (el numerador en la ecuación 3.35 tiene que ser menor o igual que el denominador). Generalmente, toma valores entre 0 y 1 pero puede ser negativo (el denominador en la ecuación 3.35 es siempre positivo, pero el numerador puede ser negativo). Valores del coeficiente próximos a 1 indican fiabilidad alta; valores próximos a 0, fiabilidad baja.

---

[4] Para comprobar la igualdad de las 3 ecuaciones, recuerde que la puntuación en el test es una combinación lineal de las puntuaciones en los ítems y, por tanto, la varianza del test puede expresarse como la suma de las varianzas y covarianzas entre ítems:

$$S_X^2 = \sum_{j}S_{X_j}^2 + \sum_{j\neq j'}S_{X_jX_l}$$

Ejemplo 3.11. Coeficiente alfa

Podemos calcular el coeficiente α con los datos del ejemplo de la Tabla 3.8. El coeficiente α, en este caso, sería:

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum S_{X_j}^2}{S_X^2}\right) = \frac{6}{5}\left(1 - \frac{0{,}233 + 0{,}267 + 0{,}233 + 0{,}233 + 0{,}233 + 0{,}267}{4}\right) = 0{,}76$$

El coeficiente α obtenido representa un valor aceptable, pues se ha obtenido con sólo 6 ítems, que nos indica que existe un grado de covariación medio-alto entre los ítems.

La cuantía del coeficiente α depende de dos factores principalmente:

1. *Consistencia interna o grado de covariación (correlación) promedio* entre los ítems. Como es lógico, un grado de covariación mayor entre dos ítems implica que el efecto de aplicar uno u otro para puntuar a las personas es menos importante. Podemos observar en la expresión [3.35] que el coeficiente α tendrá un valor mayor cuanto mayor sea el promedio de las covarianzas. Asumirá valores cercanos a cero si el promedio de las covarianzas es próximo a 0. El máximo valor de α es 1, ya que la covarianza entre dos ítems nunca puede ser mayor que las varianzas de estos (ya que el numerador nunca puede ser mayor que el denominador). El grado de covariación será mayor si los ítems están midiendo una única dimensión o rasgo (o dimensiones distintas pero correlacionadas) y mayor cuanto mejor reflejen esa dimensión (o dimensiones). Sin embargo, y para evitar malos entendidos, debemos recordar que α, por sí solo, no constituye un indicador de unidimensionalidad ya que:

   a. Se pueden estar midiendo distintas dimensiones pero correlacionadas.

   b. La covariación promedio puede llegar a ser alta incluso si un conjunto reducido de ítems no covarían con los demás

   c. Como se describe a continuación, cierto grado de multidimensionalidad del test puede compensarse incrementando el número de ítems (Cortina, 1993; Streiner, 2003). Por tanto, para concluir sobre la unidimensionalidad del test es aconsejable aplicar otras técnicas estadísticas, como el Análisis Factorial (ver capítulos 6 y 13).

2. *Número de ítems.* En la ecuación [3.35] se observa también que el coeficiente α será mayor cuanto mayor sea el número de ítems. En efecto, llamemos al promedio de las varianzas de los ítems $\overline{S}_{X_j}^2$:

$$\overline{S}^2_{X_j} = \frac{\sum_j S^2_{X_j}}{J}$$

La fórmula 3.35 se puede escribir como:

$$\alpha = \frac{\overline{S}_{X_j X_{j'}}}{\frac{J\overline{S}^2_{X_j} + J(J-1)\overline{S}_{X_j X_{j'}}}{J^2}} = \frac{J\frac{\overline{S}_{X_j X_{j'}}}{\overline{S}^2_{X_j}}}{1+(J-1)\frac{\overline{S}_{X_j X_{j'}}}{\overline{S}^2_{X_j}}}$$

Si al añadir ítems se mantiene constante el cociente entre el promedio de las covarianzas y el promedio de las varianzas, el valor del coeficiente α será mayor cuanto mayor sea *J*. Puede observarse el parecido de la estructura de esta fórmula y la de Spearman-Brown.

Ejemplo 3.12. Coeficiente alfa y unidimensionalidad del test

En las siguientes tablas (3.9, 3.10 y 3.11) se muestran las matrices de varianzas-covarianzas entre los ítems de tres pruebas; los tests *A* y *B* tienen 6 ítems, mientras que el *C* tiene 12 ítems. Los datos son ficticios para ilustrar mediante un ejemplo simple las propiedades del coeficiente α.
Para el test *A*:

$$\alpha = \frac{J}{J-1}\left(\frac{\sum_{j\neq j'} S_{X_j X_l}}{S^2_X}\right) = \frac{6}{5}\left(\frac{2{,}4}{3{,}9}\right) = 0{,}74$$

Para el test *B*:

$$\alpha = \frac{J}{J-1}\left(\frac{\sum_{j\neq j'} S_{X_j X_l}}{S^2_X}\right) = \frac{6}{5}\left(\frac{2{,}4}{3{,}9}\right) = 0{,}74$$

Para el test *C*:

$$\alpha = \frac{J}{J-1}\left(\frac{\sum_{j\neq j'} S_{X_j X_l}}{S^2_X}\right) = \frac{12}{11}\left(\frac{12}{15}\right) = 0{,}87$$

Tabla 3.9. Matriz de varianzas-covarianzas entre ítems (Test A)

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0,25	0,20	0,20	0	0	0
X_2	0,20	0,25	0,20	0	0	0
X_3	0,20	0,20	0,25	0	0	0
X_4	0	0	0	0,25	0,20	0,20
X_5	0	0	0	0,20	0,25	0,20
X_6	0	0	0	0,20	0,20	0,25

Tabla 3.10. Matriz de varianzas-covarianzas entre ítems (Test B)

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0,25	0,08	0,08	0,08	0,08	0,08
X_2	0,08	0,25	0,08	0,08	0,08	0,08
X_3	0,08	0,20	0,25	0,08	0,08	0,08
X_4	0,08	0,08	0,08	0,25	0,08	0,08
X_5	0,08	0,08	0,08	0,08	0,25	0,08
X_6	0,08	0,08	0,08	0,08	0,08	0,25

Tabla 3.11. Matriz de varianzas-covarianzas entre ítems (Test C)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	0,25	0,20	0,20	0,20	0,20	0,20	0	0	0	0	0	0
X_2	0,20	0,25	0,20	0,20	0,20	0,20	0	0	0	0	0	0
X_3	0,20	0,20	0,25	0,20	0,20	0,20	0	0	0	0	0	0
X_4	0,20	0,20	0,20	0,25	0,20	0,20	0	0	0	0	0	0
X_5	0,20	0,20	0,20	0,20	0,25	0,20	0	0	0	0	0	0
X_6	0,20	0,20	0,20	0,20	0,20	0,25	0	0	0	0	0	0
X_7	0	0	0	0	0	0	0,25	0,20	0,20	0,20	0,20	0,20
X_8	0	0	0	0	0	0	0,20	0,25	0,20	0,20	0,20	0,20
X_9	0	0	0	0	0	0	0,20	0,20	0,25	0,20	0,20	0,20
X_{10}	0	0	0	0	0	0	0,20	0,20	0,20	0,25	0,20	0,20
X_{11}	0	0	0	0	0	0	0,20	0,20	0,20	0,20	0,25	0,20
X_{12}	0	0	0	0	0	0	0,20	0,20	0,20	0,20	0,20	0,25

Podemos observar que los tests *A* y *B* tienen el mismo coeficiente α. Sin embargo, la interpretación de los resultados sería bastante distinta. Aunque los dos tests tienen el mismo número de ítems, la misma varianza y el mismo promedio para las covarianzas entre ítems, el patrón de resultados es muy distinto atendiendo a las covarianzas concretas entre ítems. En el test *A* los ítems miden dos dimensiones; los ítems del 1 al 3 miden una dimensión y los ítems del 4 al 6 miden otra dimensión. En el test *B* los 6 ítems miden una única dimensión, aunque las covarianzas entre los ítems que covarían positivamente son menores que las encontradas para el test *A*. Para el Test *C* se obtiene un coeficiente α superior (0,87). Observe que a pesar del alto valor del coeficiente obtenido, los ítems también miden dos dimensiones (los ítems del 1 al 6 miden una dimensión y los ítems del 7 al 12 miden otra dimensión). De hecho, las covarianzas entre los ítems que covarían positi-

vamente son similares a las encontradas para los ítems que covarían en el Test *A*; sin embargo, al ser el test más largo el coeficiente obtenido es mayor.

Lo anterior ilustra que la interpretación del coeficiente α debe complementarse con los resultados obtenidos a partir del análisis de ítems y del Análisis Factorial. Un coeficiente α bajo puede indicar que los diferentes ítems miden rasgos o constructos diferentes o que el test es demasiado corto.

El coeficiente α puede interpretarse como una estimación *"a la baja"* del coeficiente de fiabilidad como consistencia interna. Para interpretar el coeficiente α como un coeficiente de fiabilidad del test se requiere asumir que todos los ítems son paralelos o, al menos, esencialmente tau-equivalentes (ver capítulo 11 para la definición de tau-equivalencia). En la práctica, es muy difícil que esto se produzca. Cuando los ítems no son equivalentes, el coeficiente alfa poblacional debe interpretarse como una subestimación del coeficiente de fiabilidad como consitencia interna (Lord y Novick, 1968): $\alpha \leq \sigma^2_V / \sigma^2_X$. Esto quiere decir que si obtenemos un coeficiente α de 0,7, el coeficiente de fiabilidad podría estar, teóricamente, entre 0,7 y 1. Por tanto, en sentido estricto, el coeficiente α no puede interpretarse como un coeficiente de fiabilidad.

Una ventaja del coeficiente α es que no requiere dividir el test en distintas mitades. Cada test puede tener muchos coeficientes de fiabilidad por el método de las dos mitades pero siempre tendrá, para una muestra concreta, un único coeficiente α. Sin embargo, al calcular el coeficiente α hay que tener en cuenta una serie de precauciones. El coeficiente alfa adolece de algunos problemas comunes a los otros indicadores de consistencia interna:

1. Las respuestas a los ítems pueden correlacionar excesivamente, independientemente de su contenido, si el test es de velocidad o hay efectos de fatiga.
2. También hay que ser cauto cuando existen grupos de ítems que comparten su especificidad (p.ej., si conjuntos de ítems de un test de comprensión lectora se refieren a pasajes comunes). En ambos casos es preferible obtener otros indicadores de fiabilidad. Una solución sencilla para el último caso, puede ser construir testlets (p.ej., cada testlet sería la suma de las puntuaciones de los ítems que se refieren a un pasaje común) y calcular el coeficiente alfa tomando los testlets como ítems.
3. Debe evitarse aumentar el coeficiente α artificialmente, incluyendo ítems redundantes en el test (p.ej., ítems muy parecidos en el enunciado).
4. Finalmente, el coeficiente alfa no es sensible al efecto de los errores debidos a factores transitorios (Becker, 2000; Green, 2003; Schmidt y Hunter, 1996, 1999). En la presencia de este tipo de errores, el coeficiente α es una sobreestimación del coeficiente de fiabilidad.

El coeficiente α es probablemente el indicador de fiabilidad más utilizado (Hogan, Benjamin y Brezinski, 2000). Sin embargo, la discusión sobre su interpretación sigue generando polémica. En el número de marzo de 2009, una de las revistas psicométricas más prestigiosas, *Psychometrika*, dedicó un número especial sobre la interpretación, usos, abu-

sos y alternativas al coeficiente α como aproximación a la fiabilidad. La interpretación del coeficiente α puede ser especialmente problemática si el test no es unidimensional.

En relación al uso del test, debe distinguirse entre el valor del coeficiente alfa como un indicador de la consistencia interna o como un indicador de que el test puede ser utilizado en la práctica. Un coeficiente alfa de 0,60 puede indicar una alta consistencia interna si la prueba tiene sólo 6 ítems. Sin embargo, esa alta consistencia interna no legitima su uso, porque la precisión de nuestras medidas será claramente insuficiente.

Existen otros muchos indicadores relacionados con el coeficiente alfa, pero la mayoría de ellos raramente son aplicados en la práctica. Por ejemplo, los coeficientes *KR-21* y *KR-20* (Kuder y Richardson, 1937) son casos particulares del coeficiente α para ítems dicótomicos. Mientras que *KR-20* es matemáticamente equivalente al coeficiente alfa, el coeficiente *KR-21* no lo es, ya que en su cómputo se asume que los ítems tienen la misma dificultad y se cumplirá siempre que $KR\text{-}21 \leq \alpha$.

El error típico de medida

Definición

Asumiendo el postulado fundamental del Modelo Clásico, que expresa la relación $X = V + E$, hemos observado que se cumple la siguiente relación para datos poblacionales $\sigma^2_X = \sigma^2_V + \sigma^2_E$. También hemos demostrado que $\rho_{XX} = 1 - \sigma^2_E / \sigma^2_X$, de donde se deduce que la desviación típica de los errores puede obtenerse a partir de la expresión:

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX}} \qquad [3.36]$$

σ_E es la desviación típica de los errores al aplicar un test en la población. En el Modelo Clásico suele asumirse que σ_E expresa también la desviación típica de los errores de medida que obtendríamos, para una persona, al aplicarle distintas formas del test (σ_{Ei}). Es decir, que $\sigma_{Ei} = \sigma_E$. A la desviación típica de los errores de medida (σ_E) se denomina *error típico de medida*. Representa una medida de precisión: cuanto más cercano a 0 sea el error típico de medida de un test, eso significará que dicho test proporciona a cada persona una puntuación X cercana a su nivel de rasgo V. El error típico de medida es muy importante, ya que indica la variabilidad de las puntuaciones X si tomáramos para la misma persona distintas medidas. En efecto, para un individuo i, la variabilidad de las puntuaciones a través de distintas formas paralelas se explica por la varianza de los errores (ya que, siendo su puntuación verdadera constante a través de las formas, $\sigma^2_{Vi} = 0$):

$$\sigma^2_{X_i} = \sigma^2_{V_i} + \sigma^2_{E_i} = \sigma^2_{E_i}$$

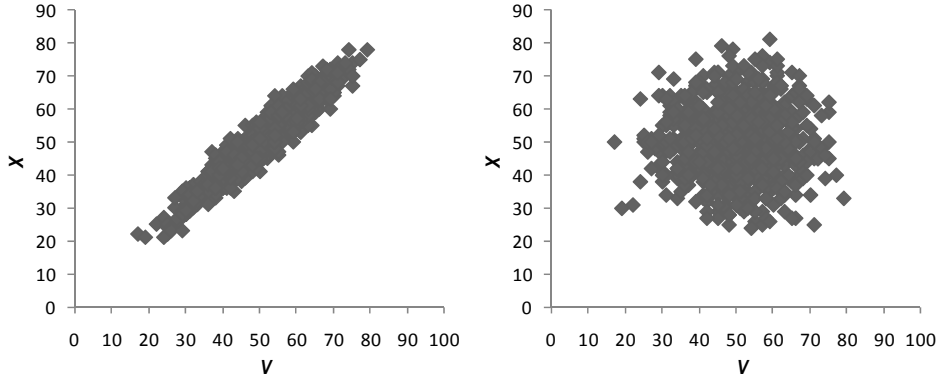
Si el error típico de medida σ_{Ei} es 0, eso quiere decir que el evaluado i obtendrá siempre la misma puntuación X en las distintas mediciones (como el test en ese caso es máximamente preciso, la puntuación X del evaluado coincidirá siempre con su puntuación V). Cuanto menos preciso sea el test, mayor será σ_{Ei}. Si el coeficiente de fiabilidad de las puntuacio-

nes fuera 0 ($\rho_{XX} = 0$), entonces el error típico de medida sería el máximo posible, $\sigma_E = \sigma_X$; esto quiere decir que cuando trabajamos con una prueba poco precisa la variabilidad de las puntuaciones observadas para una persona en distintas mediciones va a ser tan grande como la variabilidad de las puntuaciones observadas en la población. Luego el test resultará poco útil para informarnos sobre los niveles de atributo de las personas.

Ejemplo 3.13. Coeficiente de fiabilidad y error típico de medida

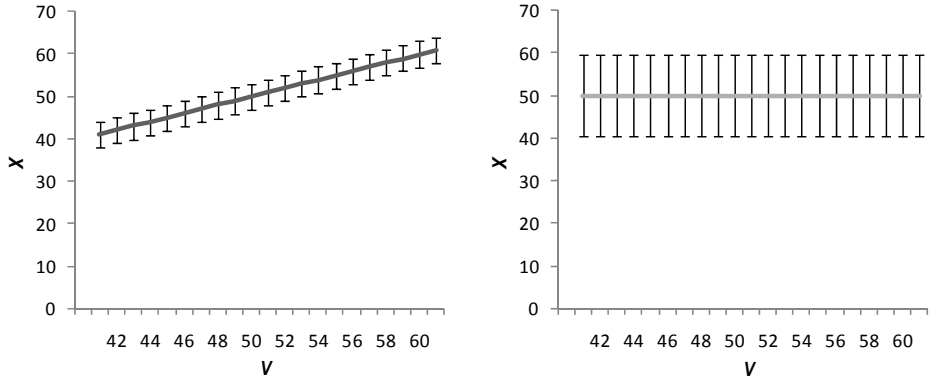
En la figura 3.2 se presentan 2 gráficos de dispersión entre *V* y *X* para dos tests. La media y desviación típica de *X* son 50 y 9,65, respectivamente. El primero representa lo que ocurre para un test con un coeficiente de fiabilidad de 0,9 y un error típico de 3,05 (relación lineal positiva y elevada). El segundo, lo que ocurre para un test con un coeficiente de fiabilidad de 0 y un error típico de 9.65 (no existe relación lineal entre *X* y *V*).

Figura 3.2. Relación entre *V* y *X* para dos tests



En la figura 3.3 se representa la puntuación media y la variabilidad en *X* como función de *V* para esos mismos tests para las puntuaciones verdaderas entre 41 y 61.

Figura 3.3. Puntuación esperada y variabilidad en *X* como función de *V* para dos tests.



Para el test con alta fiabilidad (izquierda) el valor esperado en X es función de V. Por ejemplo, las personas con una puntuación verdadera de 45 tienen una puntuación esperada en el test de 45 y sus puntuaciones suelen oscilar en la mayoría de los casos entre 42 y 48. Para el test con fiabilidad nula (derecha), el valor esperado en X no depende de V. Por ejemplo, las personas con una puntuación verdadera de 45 tienen una puntuación esperada de 50 (la media del test) y sus puntuaciones suelen oscilar aproximadamente entre 40 y 60. En ese caso, la puntuación en el test no nos informa del nivel de rasgo. Puede observarse que la amplitud de los intervalos en cada test (42-48 y 40-60) se relaciona inversamente con su fiabilidad y es proporcional al error típico de medida.

De lo anterior debe deducirse que el tamaño del error típico de medida debe interpretarse en relación a la variabilidad de las puntuaciones empíricas. Si σ_E es 1 y σ_X es 15, nuestro test será más preciso que si σ_E es 0,8 y σ_X es 1.

En una muestra concreta el error típico de medida se estima como:

$$S_E = S_X\sqrt{1 - r_{XX}} \qquad [3.37]$$

Ejemplo 3.14. Cálculo del error típico de medida

En un test la desviación típica es 2,832 y el coeficiente de fiabilidad es 0,771; el error típico de medida se obtendría como:

$$S_E = S_X\sqrt{1 - r_{XX'}} = 2{,}832\sqrt{1 - 0{,}771} = 1{,}355$$

Esto quiere decir que si aplicáramos a una persona tests paralelos, la desviación típica de las puntuaciones empíricas sería 1,355.

Aplicaciones del error típico de medida

El error típico de medida nos sirve para saber: (1) el rango de puntuaciones en el cuál se encuentra la puntuación verdadera de una persona; (2) si la diferencia de puntuaciones observadas entre dos personas expresa una diferencia en parte verdadera; (3) si el cambio en las puntuaciones observadas de una persona después de una intervención refleja un cambio en parte verdadero.

Desde el Modelo Clásico se suele asumir que la distribución de las puntuaciones de una persona en las distintas formas paralelas es normal con media su puntuación verdadera y desviación típica el error típico de medida:

$$X_i \sim N(V_i, \sigma_E) \qquad [3.38]$$

Puesto que se asume la distribución normal puede decirse que los valores de la variable estarán entre el valor $V_i - z_{1-\alpha/2}\sigma_E$ y el valor $V_i + z_{1-\alpha/2}\sigma_E$ con una probabilidad $1 - \alpha$[5], donde $z_{1-\alpha}$ es el valor z que deja por debajo una probabilidad $1 - \alpha$ en la distribución normal. Por ejemplo, con $\alpha = 0{,}05$, si $\sigma_E = 1{,}355$ y $V_i = 5$ podremos decir que los valores de X estarán, en el 95% de las mediciones, entre 2,344 (= 5 – (1,96)1,355) y 7,656 (= 5 + (1,96)1,355).

En la realidad operamos al revés, pues no conocemos V_i sino X_i y queremos establecer un intervalo de confianza sobre V_i. Además, se trabaja con la estimación muestral del error típico de medida. Para ello, se procede de la siguiente manera para establecer los límites inferior y superior del intervalo de confianza:

$$\begin{aligned} V_{Li} &= X_i - z_{1-\alpha/2} S_E \\ V_{Ls} &= X_i + z_{1-\alpha/2} S_E \end{aligned} \qquad [3.39]$$

Por ejemplo, con $\alpha = 0{,}05$, si $S_E = 1{,}355$ y $X_i = 5$, diremos que los valores de V_i estarán entre 2,344 (V_{Li} = 5 – (1,96)1,355) y 7,656 (V_{Ls} = 5 + (1,96)1,355) con un nivel de confianza del 95% (al establecer de ese modo el intervalo sobre la puntuación verdadera, nos equivocaremos en nuestra afirmación en el 5% de los casos).

Además, mediante el error de medida podemos saber si una diferencia en puntuaciones empíricas refleja una diferencia no nula de puntuaciones verdaderas. Por ejemplo, un test impreciso puede proporcionar a dos personas puntuaciones empíricas diferentes aunque sus niveles de rasgo sean iguales. Utilizando los procedimientos de las estadística inferencial, podemos contrastar con cierta probabilidad si dos puntuaciones empíricas diferentes suponen o no niveles de rasgo distintos, o si un incremento en la puntuación empírica de una persona refleja un incremento en su nivel de rasgo.

Para realizar el contraste, partimos de una situación en la que observamos una diferencia entre dos puntuaciones empíricas obtenidas en el mismo test (o en tests paralelos), X_1 y X_2, y queremos saber si la diferencia entre esas puntuaciones empíricas refleja una diferencia en los niveles de rasgo verdaderos, V_1 y V_2. Partimos de que la diferencia entre X_1 y X_2 se distribuye normalmente:

$$X_1 - X_2 \sim N(V_1 - V_2, S_E\sqrt{2}) \qquad [3.40]$$

Y esto nos permite obtener:

$$Z = \frac{(X_1 - X_2) - (V_1 - V_2)}{S_E\sqrt{2}} \sim N(0,1)$$

[5] A pesar de denominarse de la misma forma, no debe confundirse el nivel de significación α de un contraste de hipótesis (la probabilidad asociada a la zona de rechazo de H_0) con el coeficiente α de Cronbach.

Lo más usual es contrastar si la diferencia entre X_1 y X_2 es estadísticamente distinta de 0. Bajo la hipótesis nula, se considera que $V_1 - V_2 = 0$ y obtenemos el estadístico de contraste:

$$Z = \frac{(X_1 - X_2)}{S_E\sqrt{2}} \qquad [3.41]$$

El Cuadro 3.1 resume los pasos de este contraste. El contraste puede ser bilateral (p.ej., la hipótesis nula es que *no* hay diferencias en puntuaciones verdaderas) o unilateral (p.ej., la hipótesis nula es que la persona *no* ha mejorado su puntuación verdadera después del tratamiento). Si el valor *Z* se encuentra en la zona crítica, admitiremos, con la probabilidad establecida α de equivocarnos, que las puntuaciones V_1 y V_2 son distintas (o que ha habido una mejora). De lo contrario, admitiremos que, dada la precisión del test, no podemos concluir que la diferencia en puntuaciones empíricas exprese una diferencia (o mejora) en el verdadero nivel de rasgo. Como se muestra en el cuadro, también podemos establecer un intervalo de confianza para la diferencia en puntuaciones verdaderas.

Cuadro 3.1. Resumen del contraste sobre puntuaciones verdaderas

1. *Hipótesis:*
   a. Contraste bilateral: H_0: $V_1 = V_2$; H_1:$V_1 \neq V_2$
   b. Contraste unilateral derecho: H_0: $V_1 \leq V_2$; H_1: $V_1 > V_2$
   c. Contraste unilateral izquierdo: H_0: $V_1 \geq V_2$; H_1: $V_1 < V_2$
2. *Supuestos:* Se asume una distribución normal para $X_1 - X_2 \sim N(V_1 - V_2, S_E\sqrt{2})$
3. *Estadístico del contraste:*

$$Z = \frac{X_1 - X_2}{S_E\sqrt{2}}$$

4. *Distribución muestral: Z* se distribuye normalmente con media 0 y desviación típica 1.
5. *Zona crítica*
   a. Contraste bilateral: $Z \leq z_{\alpha/2}$ y $Z \geq z_{1-\alpha/2}$
   b. Contraste unilateral derecho: $Z \geq z_{1-\alpha/2}$
   c. Contraste unilateral izquierdo: $Z \leq z_{\alpha/2}$
6. *Regla de decisión:* se rechaza H_0 si el estadístico de contraste cae en la zona crítica; en caso contrario, se mantiene.
7. *Intervalo de confianza:* $IC_{V_1-V_2} = X_1 - X_2 \pm z_{1-\alpha/2}S_E\sqrt{2}$

Ejemplo 3.15. Contraste de puntuaciones verdaderas

En la escala de Neuroticismo de un test dos personas obtienen unas puntuaciones directas de 13 y 15 puntos, respectivamente. La desviación típica del test es 2,832. El investigador se pregunta si, con probabilidad 0,95, puede concluir que ambas personas difieren en el rasgo o nivel verdadero.

En este caso, los pasos a seguir serían:

1. *Hipótesis:* H_0: $V_1 = V_2$; H_1:$V_1 \neq V_2$ (contraste bilateral).
2. *Supuestos:* Se asume una distribución normal para $X_1 - X_2 \sim N(V_1 - V_2, S_E\sqrt{2})$.
3. *Estadístico del contraste:*

$$Z = \frac{X_2 - X_1}{S_E\sqrt{2}} = \frac{15-13}{1{,}355\sqrt{2}} = \frac{15-13}{1{,}916} = 1{,}044$$

4. *Distribución muestral:* Z se distribuye normalmente con media 0 y desviación típica 1.
5. *Zona crítica:* $Z \leq -1{,}96$ y $Z \geq 1{,}96$
6. *Regla de decisión:* como $-1{,}96 < 1{,}044 < 1{,}96$, se mantiene H_0.
7. *Intervalo de confianza:* $IC_{V_1-V_2} = X_1 - X_2 \pm z_{1-\alpha/2} S_E\sqrt{2} =$

$$= 2 \pm 1{,}96(1{,}916) = (-1{,}756;\ 5{,}756)$$

Con un nivel de confianza del 95%, la zona de aceptación queda establecida entre los límites $z_{0,025} = -1{,}96$ y $z_{0,975} = 1{,}96$, con lo cual, dada la precisión del test, no podemos concluir, con $\alpha = 0{,}05$, que las dos personas difieran en el verdadero nivel de rasgo. Esto es lógico, ya que la diferencia encontrada entre las puntuaciones empíricas (2 puntos) no es mucho mayor que el error típico de medida (1,355 puntos). El intervalo de confianza nos dice que, con un nivel de confianza del 95%, la diferencia verdadera se encuentra aproximadamente entre –1,756 puntos y 5,756 puntos, que es un intervalo relativamente amplio. En esta escala, las diferencias entre dos puntuaciones empezarían a ser estadísticamente significativas (con $\alpha = 0{,}05$) a partir de 3,756 ($\cong 1{,}96(1{,}916)$) puntos.

Formas de incrementar la fiabilidad de un test

Existen varias formas de incrementar la fiabilidad de un test:

1. *Aumentar el número de ítems:* Una de las maneras de incrementar la fiabilidad de un test es aumentar el número de ítems. Para estudiar el efecto de la longitud del test, puede aplicarse la fórmula de Spearman-Brown:

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}}$$

Mediante la cual puede estudiarse cómo aumentaría la fiabilidad al incrementar el número de ítems si no hay efectos de la fatiga (que producen correlaciones entre los errores de medida de los ítems), si las formas añadidas son paralelas y si los errores debidos a factores transitorios son pequeños (Feldt y Brennan, 1989; Schmidt et al., 2003). Además, debe tenerse una precaución adicional: al añadir ítems nuevos no debe buscarse el aumento artificial del coeficiente α incluyendo ítems redundantes.

2. *Eliminar ítems problemáticos.* Además de incrementar el número de ítems, pueden eliminarse los ítems problemáticos (cuya correlación con la puntuación en el resto del test es baja). Entre los ítems de igual variabilidad, los de mayor correlación biserial puntual con el test, r_{bp} , son los que más contribuyen a incrementar α ya que tendrán mayor promedio de covarianzas con el resto de los ítems. Para ítems con igual varianza, α es proporcional a los valores r_{bp} de los ítems ya que:

$$\alpha = \left(\frac{J}{J-1}\right)\left(1-\frac{\sum_{j=1}^{J} S_{X_j}^2}{\left(\sum_{j=1}^{J} S_{X_j} r_{bp,j}\right)^2}\right) \qquad [3.42]$$

Y, si las varianzas son iguales:

$$\alpha = \frac{J}{J-1}\left(1-J\Big/\left(\sum_{j=1}^{J} r_{bp,j}\right)^2\right) \qquad [3.43]$$

Si en la fase de análisis de ítems tenemos como objetivo elaborar un test con elevada consistencia interna, tenemos que quedarnos con los ítems que manifiestan una mayor correlación ítem-test. Sin embargo, esta regla, de uso frecuente, debe aplicarse con precaución ya que:

a. Si un ítem correlaciona de forma aceptable con el resto del test no debería eliminarse incluso si con ello cambia poco o aumenta la fiabilidad, ya que existen otras propiedades psicométricas del test que podrían verse afectadas (p.ej., el nivel de representación de los contenidos).

b. Si la muestra es pequeña, es probable que el aumento en el coeficiente α al quitar un ítem con baja r_{bp} no se replique en una nueva muestra.

c. Si el objetivo del estudio psicométrico no es el desarrollo de un nuevo test puede ser cuestionable la eliminación de ítems, pues ello dificultará la comparación de los coeficientes de fiabilidad que se obtendrían con la nueva versión de la prueba.

3. *Mejorar las condiciones de aplicación.* Finalmente, tras la aplicación de un test podemos detectar ciertos aspectos que se han podido descuidar (instrucciones de aplicación poco claras, tiempos de aplicación inadecuados, etc.). Al homogeneizar al máximo las condiciones de aplicación (especialmente en lo relativo a las instrucciones y a los tiempos de aplicación de la prueba) haremos que éstas no incrementen la variabilidad error en las puntuaciones.

Coeficiente de fiabilidad y características de la muestra

Actualmente se considera un error hablar de fiabilidad del test (Fan y Yin, 2003; Thompson y Vacha-Haase, 2000). Parece que es más correcto hablar de *fiabilidad de las puntuaciones* obtenidas en el test. Más que una discusión terminológica, lo que se pretende destacar es que el coeficiente de fiabilidad obtenido para un test dependerá de la muestra de personas en la cuál lo hayamos calculado (especialmente, de la variabilidad en la característica medida), de las fuentes de error a las que es sensible el coeficiente obtenido y de la situación de aplicación (p.ej., de las instrucciones proporcionadas).

La variabilidad de las puntuaciones en la muestra es uno de los factores que más puede afectar al valor del coeficiente de fiabilidad. Más concretamente, obtendremos un coeficiente de fiabilidad mayor cuanto más heterogénea (mayor varianza en el rasgo) sea la muestra. Por ejemplo, es usual que un test de Inteligencia obtenga un r_{XX} mayor en una muestra de la población general que una muestra de universitarios o en otra de personas con deficiencias cognitivas. Esto se debe a que, en último término, el coeficiente de fiabilidad es una correlación de Pearson y, por tanto, se ve afectado por los mismos factores estadísticos que ésta.

Existen fórmulas para corregir los efectos de la variabilidad, denominas como fórmulas para la *corrección del coeficiente de fiabilidad por restricción de rango*. Su aplicación no está exenta de supuestos (p.ej., que la varianza error se mantiene constante a través de los grupos) y, por tanto, de críticas. En concreto, asumiendo que la varianza de los errores es la misma en dos grupos (*A* y *B*), el coeficiente de fiabilidad en el grupo *B* puede obtenerse como:

$$\rho_{XX(B)} = 1 - \frac{\sigma^2_{X(A)}\left(1 - \rho_{XX(A)}\right)}{\sigma^2_{X(B)}} \qquad [3.44]$$

donde $\rho_{xx(A)}$ y $\rho_{xx(B)}$ indican el coeficiente de fiabilidad en los grupos *A* y *B* respectivamente; $\sigma^2_{X(A)}$ y $\sigma^2_{X(B)}$ indican las varianzas de las puntuaciones empíricas en los grupos *A* y *B*, respectivamente. Esta fórmula *no* debería aplicarse si existen razones para pensar que los grupos difieren en cuanto a la varianza de los errores (lo que puede ocurrir si la precisión del test varía mucho dependiendo del nivel de rasgo).

Ejemplo 3.16. Corrección por restricción de rango

Tras un proceso de selección se ha aplicado una prueba de Extraversión al grupo de personas seleccionadas. Se obtiene un coeficiente de fiabilidad de 0,6 y una varianza de las puntuaciones en el test de 7. El investigador se pregunta cuál habría sido el coeficiente de fiabilidad si hubiera aplicado el test en el grupo completo de aspirantes que se presentaron al proceso de selección. En el manual de la prueba se describe que su varianza es 10 en la población. Asumiendo que en el grupo de aspirantes esa sea la varianza, la estimación del coeficiente de fiabilidad para dicho grupo será:

$$\rho_{XX} = 1 - \frac{7(1-0{,}6)}{10} = 0{,}72$$

Algunos autores consideran que en un grupo de aspirantes suele haber menor variabilidad en el rasgo que en la población. Esto puede ocurrir por un efecto de autoselección (p.ej., si las personas poco extravertidas optan por no presentarse a trabajos en los que se demanda esa característica de personalidad). Estudios publicados previos pueden servir para valorar el grado en que ocurre este efecto para distintas características de personalidad y en distintos tipos de trabajos (ver por ejemplo, Ones y Viswesvaran, 2003)

Valores mínimos para los indicadores de fiabilidad

La falta de fiabilidad de las puntuaciones en un test supone que una parte importante de la variabilidad de las puntuaciones es aleatoria. Ante este problema, la pregunta podría ser: ¿A partir de qué valor del coeficiente de fiabilidad aceptamos que las puntuaciones son suficientemente fiables? La respuesta a esta pregunta es ambigüa porque está mal formulada. Es un error pensar que existen límites casi mágicos, como el 0,7, a partir de los cuales nos podemos olvidar de la falta de precisión de las puntuaciones del test (Schmidt y Hunter, 1999). Este error surge a partir de los intentos de establecer guías que nos permitan establecer valores mínimos de precisión para las pruebas. Por ejemplo, Nunnally (1967) recomendaba inicialmente valores por encima de 0,5 o 0,6 en las fases tempranas de la investigación; en versiones posteriores de su manual incrementaron el valor a 0,7 (Nunnally y Bernstein, 1994); para instrumentos que se vayan a utilizar en investigación básica recomiendan un valor mínimo de 0,8 y si se va a hacer un uso clínico 0,9 es, para ellos, el valor mínimo aceptable[6]. En realidad, el valor del coeficiente de fiabilidad que podamos aceptar debe venir fijado más por el uso específico que se vaya a hacer del test (p.ej., considerando la precisión requerida para ese uso o las consecuencias de la falta de

[6] Aunque para Streiner este último criterio es demasiado exigente y puede resultar contraproducente ya que, en su opinión, un valor tan alto sólo se puede obtener a costa de incrementar la redundancia de los ítems en el test (Streiner, 2003).

precisión) que por una regla mágica, por muy consensuada que pueda estar. Por ejemplo, sin pretender ser exhaustivos, las puntuaciones en un test pueden utilizarse:

1. *En contextos de investigación básica, para estudiar las relaciones entre constructos.* Por ejemplo, Schmidt y Hunter (1999) muestran que la correlación entre dos variables medidas con pruebas cuyo coeficiente de fiabilidad sea 0,70 se verá subestimada, en promedio, en un 30 % (p.ej., una correlación de 0,3 pasará a ser una correlación de 0,21). Esto puede hacer que relaciones reales dejen de ser estadísticamente significativas. Lo mismo ocurre si estamos comparando las puntuaciones medias de los evaluados en dos grupos (p.ej., experimental y control). Si el test no resulta suficientemente fiable, las diferencias no serán estadísticamente significativas. En este tipo de situaciones, na forma adecuada de plantearnos la pregunta es: ¿Son las puntuaciones en el test lo suficientemente fiables para detectar la relación o efecto que se pretende detectar dados los tamaños muestrales de los grupos?
2. *En contextos de selección, para elegir a los candidatos aptos para el puesto.* En estos contextos es frecuente que haya un límite en el número de plazas ofertadas. En ese caso, los requerimientos en relación a la fiabilidad del test aplicado pueden depender de su uso (es diferente aplicarlo como filtro que tomar decisiones finales), de la ratio de selección (p.ej., si se debe seleccionar al 20% con puntuaciones superiores o si se debe seleccionar al 10%), de la proporción de aspirantes aptos para el puesto y de otras consecuencias que pueda tener la aplicación.
3. *En contextos de evaluación diagnostica o de certificación, para clasificar a los evaluados en relación a varias categorías o puntos de corte.* En ese caso, la precisión requerida dependerá de los niveles de puntuaciones que deseamos discriminar y de las consecuencias que pueda tener una clasificación errónea. Cuanto más próximos sean los niveles de rasgo en los que se quiere discriminar y más graves las consecuencias de una decisión errónea, mayor será la fiabilidad requerida.

Software para la Teoría Clásica de los Tests

Los programas estadísticos de carácter general (SPSS, SAS, STATISTICA) proporcionan diversos indicadores psicométricos de fiabilidad según el Modelo Clásico. Los programas comerciales LERTAP 5 (Nelson, 2001) e ITEMAN (ASC, 1988) permiten el análisis clásico de ítems y ofrecen distinta información sobre la fiabilidad de las puntuaciones en el test. Los programas TAP (Brooks y Johanson, 2003) y CIA (Kim, 1999) son muy similares al programa ITEMAN en cuanto a su funcionalidad y son de libre distribución. El programa CLM (Lopez-Pina, 2005), también de libre distribución, proporciona un gran número de indicadores de fiabilidad. TIAPLUS, desarrollado en uno de los centros de investigación psicométrica más prestigiosos (CITO, 2006) permite también el análisis clásico de ítems y la obtención de distintos estadísticos para el estudio de la fiabilidad. En España, Renom y colaboradores (2007) han desarrollado una plataforma web (www.etest.es) de análisis psicométrico que integra distintas herramientas desarrolladas previamente por el equipo (METRIX, X-PAT, etc.).

Indicadores de fiabilidad con SPSS

Los indicadores de fiabilidad pueden obtenerse en SPSS dentro del menú **Analizar** > **Escala** > **Análisis de fiabilidad.** Para obtener el coeficiente de fiabilidad por el método de las dos mitades debe elegirse (en la pestaña correspondiente) el modelo **dos mitades**. El orden en el que se introducen las variables (i.e., los ítems) en la lista **Elementos** determina qué ítems forman cada mitad. Si el número de ítems es par, las primeras *J/2* variables formarán la primera mitad y las siguientes *J*/2 variables formarán la segunda mitad del test. Si el número de ítems es impar, las primeras *(J+1)/2* variables formarán parte de la primera mitad y las siguientes *(J* – 1*)/*2 variables formarán la segunda mitad del test. En la salida de resultados se ofrece el coeficiente $_{SB}r_{XX}$ etiquetado como **Coeficiente de Spearman-Brown (Longitud igual)**. Si el número de ítems es impar, obtendremos el coeficiente corregido **[Coeficiente de Spearman-Brown (Longitud desigual)].**

El coeficiente α puede obtenerse en SPSS eligiendo el modelo **Alfa** dentro del menú **Analizar > Escala** > **Análisis de fiabilidad.** En la salida de resultados se ofrece el coeficiente α etiquetado como *"Alfa de Cronbach"* y el coeficiente α_z etiquetado como *"Alfa de Cronbach basado en los elementos tipificados"*. Este último resulta de aplicar la fórmula tras transformar las puntuaciones a escala típica, lo que puede ser conveniente si los ítems tienen diferente formato de respuesta.

Ejemplo 3.17. Coeficiente de fiabilidad por el método de las dos mitades con SPSS

En una prueba de 11 ítems de Neuroticismo aplicada a 1569 evaluados, se obtuvieron en SPSS los resultados que aparecen en las siguientes tablas.

Tabla 3.12. Estadísticos de fiabilidad con el modelo Dos mitades en SPSS

	Correlación entre formas	0,589
Coeficiente de Spearman-Brown	Longitud igual	0,741
	Longitud desigual	0,743
	Dos mitades de Guttman	0,739

Tabla 3.13. Estadísticos descriptivos con el modelo Dos mitades en SPSS

	Media	Varianza	Desviación típica	N de elementos
Parte 1	3,78	2,793	1,671	6[a]
Parte 2	2,56	2,264	1,505	5[b]
Ambas partes	6,34	8,019	2,832	11

a. Los elementos son: u1, u2, u3, u4, u5, u6.

b. Los elementos son: u7, u8, u9, u10, u11.

Tabla 3.14. Estadísticos de fiabilidad con el modelo Alfa en SPSS

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
0,771	0,773	11

En este caso, ${}_{SB}r_{XX}$ es igual a 0,741 y el valor corregido, 0,743. Como puede observarse, los valores son bastante parecidos. Concluiríamos que el 74% de la varianza del test se debe a la varianza verdadera en el nivel de rasgo. En este caso el coeficiente α es 0,771, lo que indica que el grado de consistencia interna (o covariación media entre los ítems) es medio-alto. Además, el coeficiente α es mayor que el coeficiente de fiabilidad por el método de las dos mitades. Esto quiere decir que probablemente existen otras formas de dividir el test en dos mitades que dan lugar a mayores coeficientes de fiabilidad.

Apéndice

Otras consideraciones sobre el concepto de puntuación verdadera

A lo largo del capítulo hemos ofrecido una definición operacional de puntuación verdadera, según la cual se considera como el promedio (valor esperado) de las puntuaciones observadas que obtendría una persona en un número elevado de aplicaciones: $V_i = \varepsilon_f(X_{ij})$. Es importante ser consciente de que a partir de esta definición se establece que la puntuación verdadera depende no sólo de la *persona* sino del *instrumento* utilizado y de las *condiciones* de aplicación. Por tanto, la puntuación verdadera de una persona dependerá de su nivel de rasgo, de las propiedades del instrumento de medición (dificultad de los ítems, longitud del test, etc) y de las condiciones de aplicación (p.ej., en qué grado las instrucciones le alientan a responder al azar cuando desconoce la respuesta). Si el test fuera una prueba de conocimientos de 30 ítems, que se aplica informando a los evaluados que se les van a penalizar los errores, la puntuación verdadera de una persona es la puntuación promedio que obtendría esa persona en distintas pruebas de conocimientos de 30 ítems con las mismas especificaciones de contenido, dificultad e instrucciones de aplicación (p.ej., en relación a la penalización de los errores).

El tema es más complejo de lo que parece a primera vista. Si una característica de la aplicación (p.ej., tiempo de aplicación de la prueba) se mantiene constante a través del proceso de medición, su efecto en las puntuaciones observadas pasará automáticamente a formar parte de la puntuación verdadera (en ese caso, una puntuación verdadera *específica* que reflejaría el valor del atributo en el contexto concreto de aplicación). Por otro lado, si la misma característica de la aplicación no está controlada por el aplicador y fluctúa de una ocasión a otra, la puntuación verdadera (que podríamos denominar *genérica)* reflejaría un efecto promedio de la variable de aplicación y la variabilidad en las puntuaciones observadas provocada por la variabilidad en las condiciones de aplicación pasaría a formar parte del error.

Todo lo anterior implica que la puntuación verdadera no tiene por qué representar el nivel verdadero en el atributo que pretendemos medir; V es, simplemente, un promedio de lo que la persona obtendría en tests "como el nuestro". El coeficiente de fiabilidad sólo nos informa de cómo variaría la puntuación X de la persona en distintas aplicaciones. Desde luego, un coeficiente de fiabilidad bajo indica que estamos midiendo un atributo de forma imprecisa, pero un coeficiente de fiabilidad al-

to no implica necesariamente que estemos midiendo el atributo que queremos medir. Esto último es una cuestión de validez de las puntuaciones, a la que se dedicará el capítulo 5.

Intervalos de confianza para los estimadores de fiabilidad

Cada vez es más frecuente que para cualquier índice estadístico se exija informar del intervalo confidencial, que nos indica entre qué valores puede encontrarse el estadístico en la población. Por ejemplo, un valor r_{XX} = 0,7 es poco informativo si se ha obtenido en una muestra de 20 personas. Fan y Thompson (2001) resumen los procedimientos más usuales para construir intervalos de confianza.

Coeficiente α

En este apartado utilizaremos el símbolo $\hat{\alpha}$ para referirnos al valor del coeficiente obtenido en la muestra y el símbolo α para referirnos al valor del coeficiente obtenido en la población. Para el coeficiente α, el intervalo de confianza puede obtenerse de la forma:

$$IC_{\inf}(\alpha) = 1 - \left((1-\hat{\alpha})F_{\alpha/2,gl1,gl2}\right) \qquad IC_{\sup}(\alpha) = 1 - \left((1-\hat{\alpha})F_{1-\alpha/2,gl1,gl2}\right)$$

donde $IC_{\inf}(\alpha)$ y $IC_{\sup}(\alpha)$ son los límites inferior y superior del intervalo respectivamente; $\hat{\alpha}$ es la estimación muestral de α; F representa los valores de la distribución F para los valores de probabilidad acumulada $\alpha/2$ y $1 - \alpha/2$, con grados de libertad $gl_1 = (N - 1)$ y $gl_2 = (N - 1)(J - 1)$.

Coeficientes de fiabilidad como correlación entre formas paralelas o coeficiente de fiabilidad test-restest

Para coeficientes de fiabilidad que pueden interpretarse directamente como correlaciones (coeficiente de fiabilidad test-retest, coeficiente de fiabilidad como correlación entre formas paralelas) se pueden aplicar los procedimientos estadísticos usuales cuando se trabaja con correlaciones; los límites del intervalo confidencial se pueden obtener mediantes los siguientes pasos:

1. Se transforma el coeficiente de fiabilidad, mediante una transformación Z de Fisher:

$$z_{r_{XX}} = 0{,}5 \ln\left(\frac{1 + r_{XX}}{1 - r_{XX}}\right)$$

2. Se calcula la desviación típica de la distribución muestral de z_{rxx}:

$$\sigma_{z_{rXX}} = \frac{1}{\sqrt{N-3}}$$

3. Se obtienen los intervalos de confianza para z_{rxx}:

$$IC_{\inf}(z_{r_{XX}}) = z_{r_{XX}} - z_{1-\alpha/2}\sigma_{z_{r_{XX}}} \qquad IC_{\sup}(z_{r_{XX}}) = z_{r_{XX}} + z_{1-\alpha/2}\sigma_{z_{r_{XX}}}$$

4. Se transforman los límites del intervalo a la escala de correlaciones, mediante una transformación inversa *Z* de Fisher:

$$IC_{\inf}(r_{XX}) = \frac{\left(\exp\left(2IC_{\inf}\left(z_{r_{XX}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\inf}\left(z_{r_{XX}}\right)\right)+1\right)} \qquad IC_{\sup}(r_{XX}) = \frac{\left(\exp\left(2IC_{\sup}\left(z_{r_{XX}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\sup}\left(z_{r_{XX}}\right)\right)+1\right)}$$

Coeficiente de fiabilidad por el método de las dos mitades

Para el coficiente de fiabilidad por el método de las dos mitades se puede obtener el intervalo de confianza de una forma similar. Recuerde que el punto de partida es la correlación entre las dos mitades, $r_{X_I X_P}$. Podemos calcular los intervalos de confianza para la correlación $r_{X_I X_P}$, siguiendo el procedimiento anterior:

$$IC_{\inf}(r_{X_I X_P}) = \frac{\left(\exp\left(2IC_{\inf}\left(z_{r_{X_I X_P}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\inf}\left(z_{r_{X_I X_P}}\right)\right)+1\right)} \qquad IC_{\sup}(r_{X_I X_P}) = \frac{\left(\exp\left(2IC_{\sup}\left(z_{r_{X_I X_P}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\sup}\left(z_{r_{X_I X_P}}\right)\right)+1\right)}$$

y aplicar la formula de Spearman-Brown para obtener los límites del intervalo:

$$IC_{\inf}(r_{XX}) = \frac{2IC_{\inf}(r_{X_I X_P})}{1 + IC_{\inf}(r_{X_I X_P})} \qquad IC_{\sup}(r_{XX}) = \frac{2IC_{\sup}(r_{X_I X_P})}{1 + IC_{\sup}(r_{X_I X_P})}$$

Ejemplo 3.18. Intervalos de confianza para los coeficientes de fiabilidad

Coeficiente alfa

En una muestra de 1569 personas y un test de 11 ítems, se obtuvo un $\hat{\alpha}$ = 0,771; en ese caso, los límites del intervalo de confianza (con un nivel de confianza del 95%) para el coeficiente α son:

$$IC_{\inf}(\alpha) = 1 - \left((1-\hat{\alpha})F_{\alpha/2,gl1,gl2}\right) = 1 - \left((1-0{,}711)1{,}075\right) = 0{,}754$$
$$IC_{\sup}(\alpha) = 1 - \left((1-\hat{\alpha})F_{1-\alpha/2,gl1,gl2}\right) = 1 - \left((1-0{,}771)0{,}928\right) = 0{,}787$$

Lo que indica que podemos afirmar, con una confianza del 95%, que el coeficiente α en la población se encuentra entre 0,754 y 0,787. En este caso, el intervalo es estrecho porque el tamaño de la muestra es grande (*N* = 1569). También puede obtenerse el intervalo de confianza con SPSS. En el menú **Escalas** > A**nálisis de fiabilidad**, se selecciona el modelo **Alfa**; en **Estadísticos**, se selecciona **Coeficiente de correlación intraclase** (Modelo: **Dos factores, efectos mixtos**; Tipo: **consistencia**) y se obtiene una tabla similar a la 3.15.

Tabla 3.15. Coeficiente de correlación intraclase

		Intervalo de confianza 95%	
	Correlación intraclase	Límite inferior	Límite superior
Medidas promedio	0,771	0,754	0,787

Coeficiente de fiabilidad como correlación entre formas paralelas

En la misma muestra, la correlación entre el test de 11 ítems y otra forma paralela es también r_{xx} = 0,771. Los intervalos de confianza, con un nivel de confianza del 95%, se obtendrían siguiendo los pasos previamente mostrados:

1. Transformación de r_{xx} a *Z* de Fisher:

$$z_{r_{XX}} = 0{,}5 \ln\left(\frac{1+r_{XX}}{1-r_{XX}}\right) = 0{,}5n\left(\frac{1+0{,}771}{1-0{,}771}\right) = 1{,}023$$

2. Se obtienen los intervalos de confianza para z_{rxx} :

$$IC_{\inf}(z_{r_{XX}}) = z_{r_{XX}} - z_{1-\alpha/2}\sigma_{z_{r_{XX}}} = 1{,}023 - 1{,}96\frac{1}{\sqrt{1566}} = 0{,}973$$

$$IC_{\sup}(z_{r_{XX}}) = z_{r_{XX}} + z_{1-\alpha/2}\sigma_{z_{r_{XX}}} = 1{,}023 + 1{,}96\frac{1}{\sqrt{1566}} = 1{,}072$$

4. Se aplica la transformación inversa *Z* de Fisher:

$$IC_{\inf}(r_{XX}) = \frac{\left(\exp\left(2IC_{\inf}\left(z_{r_{XX}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\inf}\left(z_{r_{XX}}\right)\right)+1\right)} = \frac{\left(\exp\left(2(0{,}973)\right)-1\right)}{\left(\exp\left(2(0{,}973)\right)+1\right)} = 0{,}750$$

$$IC_{\sup}(r_{XX}) = \frac{\left(\exp\left(2IC_{\sup}\left(z_{r_{XX}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\sup}\left(z_{r_{XX}}\right)\right)+1\right)} = \frac{\left(\exp\left(2(1{,}072)\right)-1\right)}{\left(\exp\left(2(1{,}072)\right)+1\right)} = 0{,}790$$

Lo que indicaría que podemos afirmar, con una probabilidad 0,05 de equivocarnos, que el coeficiente de fiabilidad en la población estará entre 0,75 y 0,79.

Coeficiente de fiabilidad por el método de las dos mitades

Obtenemos, para los mismos datos que la correlación entre formas es 0,589 y el coeficiente de fiabilidad por el método de las dos mitades es 0,741. Los intervalos de confianza pueden obtenerse realizando los siguiente cálculos:

1. Transformación de $r_{X_I X_P}$ a *Z* de Fisher:

$$z_{r_{X_I X_P}} = 0{,}5n\left(\frac{1+r_{r_{X_I X_P}}}{1-r_{r_{X_I X_P}}}\right) = 0{,}5\ln\left(\frac{1+0{,}589}{1-0{,}589}\right) = 0{,}676$$

2. Se obtienen los intervalos de confianza para la *Z* de Fisher:

$$IC_{\inf}(z_{r_{X_I X_P}}) = z_{r_{X_I X_P}} - z_{1-\alpha/2}\sigma_{z_{r_{X_I X_P}}} = 0{,}676 - 1{,}96\frac{1}{\sqrt{1566}} = 0{,}627$$

$$IC_{\sup}(z_{r_{X_I X_P}}) = z_{r_{X_I X_P}} + z_{1-\alpha/2}\sigma_{z_{r_{X_I X_P}}} = 0{,}676 + 1{,}96\frac{1}{\sqrt{1566}} = 0{,}726$$

3. Se aplica la transformación inversa *Z* de Fisher:

$$IC_{\inf}(r_{X_I X_P}) = \frac{\left(\exp\left(2IC_{\inf}\left(z_{r_{X_I X_P}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\inf}\left(z_{r_{X_I X_P}}\right)\right)+1\right)} = \frac{\left(\exp(2(0{,}627))-1\right)}{\left(\exp(2(0{,}627))+1\right)} = 0{,}556$$

$$IC_{\sup}(r_{X_I X_P}) = \frac{\left(\exp\left(2IC_{\sup}\left(z_{r_{X_I X_P}}\right)\right)-1\right)}{\left(\exp\left(2IC_{\sup}\left(z_{r_{X_I X_P}}\right)\right)+1\right)} = \frac{\left(\exp(2(0{,}726))-1\right)}{\left(\exp(2(0{,}726))+1\right)} = 0{,}620$$

4. Se obtienen los intervalos:

$$IC_{\inf}(r_{XX}) = \frac{2IC_{\inf}(r_{X_I X_P})}{1+IC_{\inf}(r_{X_I X_P})} = \frac{2(0{,}556)}{1+0{,}556} = 0{,}715$$

$$IC_{\sup}(r_{XX}) = \frac{2IC_{\sup}(r_{X_I X_P})}{1+IC_{\sup}(r_{X_I X_P})} = \frac{2(0{,}620)}{1+0{,}620} = 0{,}765$$

Lo que indicaría que podemos afirmar, con una probabilidad 0,05 de equivocarnos, que el coeficiente de fiabilidad en la población estará entre 0,715 y 0,765.