6

Crítica de la Teoría Clásica de la Fiabilidad

1. Factores que afectan la fiabilidad del test

Contrariamente a lo que podría esperarse un test no tiene *un* coeficiente de fiabilidad, sino muchos. Tantos como métodos distintos, bajo condiciones distintas o en distintas muestras estimemos.

El coeficiente de fiabilidad de un test que obtengamos depende de diversos factores:

1. *El método de estimación de la fiabilidad que utilicemos.*

Para un mismo instrumento y para una misma muestra cada uno de los métodos (formas paralelas, test-retest, etc.) arroja valores diferentes.

El método de las formas paralelas refleja el grado de relación lineal entre formas paralelas (mezclado indisolublemente con otros factores espurios).

El método test-retest refleja el grado de estabilidad temporal de las puntuaciones (mezclado indisolublemente con otros factores espurios).

El método de las mitades refleja la relación lineal entre las mitades (mezclado indisolublemente con otros factores espurios).

2. Las condiciones concretas seleccionadas para aplicar el método.

En la práctica, si disponemos de tres o más formas paralelas y las aplicamos sobre una misma muestra, las correlaciones obtenidas entre cada par diferirán. También diferirán los resultados si disponiendo de dos formas paralelas variamos el lapso de tiempo entre ellas.

En el método test-retest los resultados también variarán en función del lapso de tiempo entre test y retest.

En el método de las dos mitades los resultados variarán en función de qué partición del test efectuemos para obtener las mitades.

Cualquier otro tipo de variación ambiental indeseada que se produzca puede, teóricamente, introducir efectos espurios sobre el coeficiente de fiabilidad en cada uno de los métodos.

3. Las características y tamaño de la muestra.

Un mismo test estimada su fiabilidad por un mismo método en dos muestras distintas de la misma población producirá coeficientes de fiabilidad distintos.

En general el error muestral depende del tamaño de la muestra. Cuanto mayor es la muestra menor error muestral puede esperarse. Considerado como un coeficiente de correlación pueden aplicarse al coeficiente de fiabilidad las herramientas estadísticas inferenciales que se aplican al coeficiente de correlación de Pearson.

Si se estima el coeficiente de fiabilidad en submuestras caracterizadas por alguna razón diferencial, por ejemplo, submuestras de distinto nivel de edad, esta variación entre muestras puede ser mayor. Si las submuestras difieren realmente de algún modo que pueda ser relevante para el test pueden aparecer diferencias en las distribuciones de puntuaciones, y esto puede afectar al coeficiente de fiabilidad.

Una razón esencial para esta variación es el *grado de homogeneidad o heterogeneidad* de la muestra.

El coeficiente de correlación es sensible a la variabilidad de las muestras en que se calcula. En general, cuanto mayor es la dispersión de la muestra en una o ambas variables bajo análisis mayor es el coeficiente de correlación. Por esta razón muestras más homogéneas (con menor variabilidad) tenderán a producir coeficientes de fiabilidad menores. El grado de homogeneidad de la muestra incide directamente sobre el coeficiente de fiabilidad. Esto implica que, en general, será más difícil obtener fiabilidades elevadas para tests muy específicos cuya población objetivo sea muy restringida, por ejemplo, tests destinados para un punto evolutivo preciso en sujetos caracterizados por determinado nivel de aptitudes.

4. La longitud del test.

La longitud de un test, expresada como el número de items que presenta, incide en la fiabilidad. En general cuanto más largo es un test mayor es su

fiabilidad. Esta relación se analiza con detalle al tratar la fórmula de Spearman-Brown.

Esto podría verse como una dificultad para obtener versiones cortas de las pruebas que resulten adecuadas para la medición.

Un coeficiente de fiabilidad concreto es, por tanto, función de diversas variables y grupos de variables *además* de la consistencia, estabilidad o equivalencia del instrumento.

$$r_{XX} = f(Me, Co, Mu, Lo, Fi)$$

Es decir, cualquier coeficiente de fiabilidad es función de el método de estimación (Me), las condiciones concretas de uso de ese método (Co), las características y tamaño de la muestra (Mu), la longitud del test (Lo) y la "fiabilidad" del test, definida como el grado en que el instrumento mismo no introduce variación en el resultado (sea esta fiabilidad una estabilidad, consistencia o equivalencia entre formas).

De los cinco grupos de factores de los que depende el coeficiente de fiabilidad cuatro pueden considerarse ajenos al concepto mismo de fiabilidad (Me, Co, Mu y Lo).

De los cuatro factores ajenos al concepto de fiabilidad mismo, tres de ellos están formados por grupos de variables (Me, Co, Mu) y uno, la longitud del test (Lo) es un sólo factor definido como número de items del test y cuya relación con la fiabilidad *bajo determinados supuestos* es bien conocida.

Para el factor longitud (Lo), existen fórmulas matemáticas que permiten estimar la incidencia de la longitud sobre la fiabilidad bajo determinados supuestos.

Para los otros factores, está establecida de modo explícito y formal la relación de algunas variables con la fiabilidad bajo determinados supuestos, pero no de los factores en su conjunto. En concreto está establecida explícita y formalmente la relación entre homogeneidad de la muestra y fiabilidad *bajo determinados supuestos.*

Sin embargo, la mayoría de variables a que se refieren esos factores no presentan una relación con la fiabilidad que haya sido definida de modo explícito y formal. Esto equivale a decir que no se sabe con exactitud el efecto de los mismos sobre la fiabilidad.

Una conclusión importante de este análisis es que debemos evitar hablar del "coeficiente de fiabilidad" de un test como si sólo dependiera del test y como si fuera una sola cosa.

Si se necesita estimar "coeficientes de fiabilidad" habría que especificar claramente las condiciones de su obtención

(método, circunstancias concretas de aplicación y muestra) para mejorar su interpretabilidad.

> Pueden diseñarse estudios que pongan a prueba empíricamente el efecto de determinadas variables de los grupos de factores anteriores sobre determinados coeficientes de fiabilidad. El uso de diseños factoriales más o menos complejos y del análisis de la varianza como técnica para esclarecer los efectos de esos diversos factores sobre fiabilidad excede el propósito de este texto. En esa dirección ha trabajado la *teoría de la generalizabilidad* desarrollada por Cronbach, Gleser y Rajaratnam a partir de ideas seminales y aplicaciones desarrollados por Hoyt, Lindquist y otros. Una pequeña exposición puede encontrarse en el capítulo 8 de Crocker y Algina (1986), y, en castellano, en el capítulo 8 de Santisteban (1990).

Otra consecuencia práctica es que, si se necesita estimar coeficientes de fiabilidad, estos han de estimarse en más de una muestra, mediante replicaciones directas o sistemáticas del estudio original para adquirir una cierta confianza en que los resultados obtenidos van más allá del modo en que se han dispuesto las circunstancias concretas de una estimación determinada.

2. Interpretación del coeficiente de fiabilidad

Dado que, como acabamos de ver, para un mismo test obtendremos diferentes coeficientes de fiabilidad en función de diferentes circunstancias, estas habrán de ser tenidas en cuenta en la interpretación del coeficiente de fiabilidad.

No podemos interpretar adecuadamente un coeficiente de fiabilidad sin conocer qué método se ha utilizado, en qué muestra se ha calculado, bajo qué circunstancias y por supuesto con qué versión de un test. En la interpretación de un coeficiente de fiabilidad hay que tener en cuenta pues, expresamente, los cuatro factores anteriores, además de algunas consideraciones adicionales.

En síntesis la interpretación del coeficiente de fiabilidad depende de:

> 1) El método de obtención del coeficiente de fiabilidad utilizado y las circunstancias concretas de su aplicación.
>
> 2) La muestra o muestras empleadas y en especial su tamaño, homogeneidad o heterogeneidad en la variable de interés.
>
> 3) La versión de la prueba utilizada, y en especial, la longitud del test.

Pero, además hay otros factores muy importantes a considerar en la interpretación de un coeficiente de fiabilidad:

4) El tipo de contenido psicológico que el test mide. Existen diferencias en los coeficientes de fiabilidad que se obtienen en diversos campos de medición psicológica. Así, es frecuente que de los tests aptitudes, habilidades y rendimientos se esperen valores más altos de fiabilidad que en otros campos como personalidad o actitudes.

5) Los propósitos del constructor del test, en especial en lo relativo a sensibilidad al cambio para test-retest, y homogeneidad-heterogeneidad del constructo en lo relativo a comparaciones entre mitades. Para algunas finalidades prácticas, como la predicción, selección o clasificación de sujetos, el constructor o el usuario de la prueba pueden tener exigencias especiales acerca de algún tipo de fiabilidad particular.

6) Los resultados previos de investigación en ese campo psicológico para ese tipo de instrumentos. En la práctica este puede ser un punto decisivo para interpretar adecuadamente un coeficiente de fiabilidad. Estos resultados constituyen un punto de comparación específico que ayuda a valorar el comportamiento de un test concreto.

Se está sugiriendo una interpretación comparativa que tenga en cuenta los aspectos cualitativos reseñados.

Como puede apreciarse hay que tener en cuenta muchas cosas a la vez para "leer" el significado de un coeficiente de fiabilidad. Con estas consideraciones no se puede, por ejemplo, despreciar automáticamente un test desde el punto de vista de la fiabilidad porque en un ensayo muestral para un método y unas condiciones determinadas haya obtenido un resultado pobre, pongamos de 0'4. El conocimiento de todas estas circunstancias y factores quizás pueda aconsejar modificaciones en el test o en el estudio, si es que obtener un coeficiente de fiabilidad más alto es un objetivo.

3. Procedimientos inadecuados para aumentar la fiabilidad de un test

Hay que prevenir contra ciertas *formas no aceptables de incremento de la fiabilidad.* Estas formas incluyen diversas manipulaciones del contenido o la forma de administración de la prueba que probablemente producirán ganancias en los coeficientes de fiabilidad. Veamos algunas de estas formas, en principio no recomendables, de aumentar el coeficiente de fiabilidad del test.

Duplicación semántica de items

Una forma en general no aceptable de incrementar la fiabilidad, -en especial pero no sólo la de tipo consistencia interna,- consiste en introducir en la prueba items cuyo contenido es muy semejante o igual al de otros items de la prueba.

Por ejemplo, un test de personalidad, que pretende medir aspectos relativos a la sociabilidad y modo característico de relación del sujeto, para ello pregunta:

"¿Te gusta asistir a fiestas sociales?"

Hasta aquí todo parece razonable. Pero unos cuantos items después vuelve a preguntar:

"¿Prefieres asistir a una fiesta social antes que quedarte en casa?"

Y unos cuantos items más allá, para perplejidad del sujeto que de buena fe contesta la prueba, se insiste:

"¿Disfrutas con las fiestas sociales?"

(Por cierto que después del tercer o cuarto ítem prácticamente igual el sujeto que contesta la prueba -probablemente con más sentido común que el psicólogo que la construyó- comienza a preguntarse si le falla la memoria, está sufriendo un fenómeno psicológico del tipo *"Dejà vu"* o aquello tiene algún ignoto truco en el que se le quiere hacer caer. Lo que a su vez puede llevarle a hacer una disección sintáctica y semántica cuidadosa de los items, a la busca de la diferencia perdida, y quizás a desgastar un poco su salud mental enfrentado a la absurda situación de contestar una y otra vez lo mismo.)

Estos tres items desde luego han de ser consistentes entre sí, si es que cada sujeto es mínimamente consistente en sus respuestas. Si se añaden siete items más hechos con el mismo molde, seguro que disponemos de una prueba de diez items fiable.

En primer lugar esa hipotética prueba tendrá probablemente una considerable consistencia interna. Pero además si se aplica el mismo test después de unos días también es probable que la estabilidad sea alta. Y, en tercer lugar, si construimos otros diez items más, todos ellos variantes formales (que no semánticas) de éstos, seguro que la correlación entre formas paralelas es también alta.

En estas condiciones, que, paradójicamente, van a cumplir más estrictamente los supuestos de paralelidad, hubiera bastado con un ítem para tener la información relevante. El resto es un ejercicio de redundancia y de pérdida de tiempo sin sentido. Quizás se asemeja a un intento de medir N veces una misma dimensión de un mismo objeto que no varía. Sin embargo, debido a que no se trata de un objeto sino de una persona, el intento no puede sostenerse. Esto no puede ser una medición psicológica aceptable por más que todos los coeficientes de fiabilidad pudieran llegar a ser 1.

El lector puede pensar que estamos exagerando si no se molesta en leer cuidadosamente los enunciados de algunos factores de tests de personalidad (por ejemplo) muy utilizados. El recurso de introducir algunos items que son variación formal de otros items está extendido en determinadas pruebas. Posiblemente este recurso ayudó a los constructores de la prueba a compensar el número de items entre diversas dimensiones o factores medidos, a incrementar la fiabilidad y a obtener análisis factoriales con estructuras más simples y bien formadas. En mi opinión posiblemente ninguna de estas tres cosas tenga demasiada importancia real para la calidad de una medición psicológica.

En los tests formados por items con respuesta verdadera esta redundancia puede ser más justificable, en la medida en que, por ejemplo, dos analogías o dos problemas matemáticos formalmente iguales no son semánticamente iguales, y en la medida en que se puede estar interesado en obtener información acerca de como el sujeto resuelve series de problemas del mismo tipo. Sin embargo, en pruebas de autoinforme, personalidad, actitudes, etc. creo que sólo es aceptable efectuar series de preguntas de esta índole si cada una de ellas aporta realmente algo nuevo. Es decir, si son semánticamente diferentes, si añaden nueva información cuya colección resulta útil al psicólogo.

Reducción del tiempo entre administraciones sucesivas del test

Si reducimos el tiempo entre la aplicación de una misma forma dos veces (test-retest), o entre la aplicación de dos formas paralelas, esto, generalmente, conducirá a un incremento de los coeficientes de fiabilidad correspondientes. Un caso extremo de reducción es aplicar el test y el retest el mismo día. Para el caso de formas paralelas ya no está tan claro si es legítimo o no aplicarlas sucesivamente, en la medida en que no se desee un efecto temporal test-retest. La recomendación clásica es aplicarlas con unos días de diferencia.

La memoria de las respuestas, la ausencia de la "variación debida al día a día" de la que habla Gulliksen (1950), el deseo del sujeto de mostrarse consistente con lo que acaba de decir o contestar, etc. contribuirán a abultar los coeficientes de fiabilidad.

Incremento artificial de la heterogeneidad de la muestra

Dado que cuanto más heterogénea es la muestra mayor tiende a ser el coeficiente de fiabilidad, entonces, si se forma la muestra con grupos extremos o se introducen intencionadamente sujetos pertenecientes a grupos extremos en la variable bajo medición, la fiabilidad tenderá a aumentar.

La muestra sobre la que se contrasta el instrumento es muy importante. Debe cumplir una condición: ser auténticamente una muestra adecuada de los sujetos a los que después se va aplicar el test.

Estos efectos resultan curiosos porque puede argumentarse que el tiempo entre administraciones o la heterogeneidad de la muestra no pueden alterar la fiabilidad del instrumento mismo (especialmente si se piensa en la fiabilidad como el grado en que el instrumento mismo no introduce error de medida). Sin embargo, es obvio que la reducción del tiempo y el aumento de la heterogeneidad actuarán a favor de un mayor coeficiente de fiabilidad.

En términos críticos: Si la fiabilidad del test depende de factores ajenos al test ¿cómo puede sostenerse que es la fiabilidad *del* test? Puede argumentarse que la fiabilidad depende exclusivamente del test, pero que la estimación concreta de la fiabilidad depende también de factores ajenos al test. En este caso ¿cómo podrá estimarse separada de los efectos de esos factores

4. Factores que afectan a la segunda medición

1. Todos los métodos de estimación del coeficiente de fiabilidad calculan éste como la correlación entre mediciones paralelas (se trate de partes del test, de dos administraciones del test o de dos formas del test).

2. Si las dos mediciones son paralelas (criterio estadístico) y efectivamente cumplen exactamente que sus puntuaciones verdaderas son iguales sujeto a sujeto y que su varianza de error es igual, entonces el coeficiente de correlación entre ellas es el coeficiente de fiabilidad y no hay motivo para preocuparse de ningún factor que pueda haber afectado la segunda medición. El cumplimiento estricto de las condiciones de paralelidad protege contra la duda acerca del efecto de posibles factores que puedan afectar la segunda medición.

Ahora bien, *de hecho*, el cumplimiento de las condiciones de paralelidad es usualmente relativo y, en realidad, se calculan coeficientes de correlación entre formas nominalmente paralelas.

3. El que una correlación entre mediciones al menos nominalmente paralelas pueda tratarse de modo más o menos aceptable como si fuese un coeficiente de fiabilidad, depende razonablemente de que la segunda medición se aplique sin cambios en las condiciones de administración ni en los sujetos.

Si las condiciones o los sujetos han cambiado de la primera a la segunda medición, entonces la correlación entre ellas será afectada por ese cambio y no podrá interpretarse razonablemente como si se tratará de un coeficiente de fiabilidad del instrumento.

> Por ejemplo, no admitiríamos como coeficiente de fiabilidad la correlación con un retest administrado en condiciones totalmente distintas al test. Por ejemplo, en psicología de la educación, si el retest se administra después de que los sujetos acaban de realizar un fuerte ejercicio atlético debido a una evaluación escolar de educación física. O si el retest se administra seis meses después en una muestra de niños y en una variable sujeta a fuertes cambio madurativos o, en general, evolutivos.

4. Hay factores que pueden afectar el resultado introduciendo variación en la segunda medición (bien sea esta la forma B, el retest, la segunda parte o los items impares).

Hablamos de factores que afectan la *segunda* medición no porque no haya factores que puedan afectar la primera medición sino porque, *dada una primera medición* realizada en condiciones adecuadas, se trataría de obtener una segunda sin factores que la diferenciasen de la primera.

La cuestión es que la segunda medición ha de realizarse en *condiciones constantes* respecto a las de la primera.

Esas condiciones constantes implican ausencia de cambio en los sujetos y ausencia de cambio en las condiciones de administración.

A continuación se enumeran algunos de estos factores que pueden afectar a la segunda medición.

Se han clasificado en aquellos que pueden introducir cambio en los sujetos y aquellos que pueden introducir cambio en las condiciones de administración. En realidad, algunos de ellos pueden ser vistos simultáneamente, por las mismas o por distintas razones, en ambas categorías.

Factores que pueden introducir cambios en los sujetos	Factores que pueden introducir cambios en las condiciones de administración.
-Maduración. -Aprendizaje e influencia general debida al medio social. -Actividad anterior a la administración de la prueba. -Factores que influencian el estado de ánimo de los sujetos. -Cansancio debido a otras actividades. -Estado de salud de los sujetos. - Fatiga debida a la primera prueba -Memoria de la primera prueba. -Aprendizaje debido a la primera prueba. -Conocimiento de los resultados de la primera prueba.	-El administrador de la prueba. -El local y sus condiciones ambientales. -La hora del día. -El día de la semana. -Sucesos no previstos durante la administración de las pruebas. -Pequeños errores o variaciones en las instrucciones o en los tiempos límite.

Además hay *otros efectos* que no deben despreciarse que no pueden agruparse fácilmente en estas dos categorías:

-La *mortalidad experimental*, o pérdida de sujetos entre la primera y segunda medición por las razones que sea.

-El fenómeno de ***regresión a la media***, por el cual una persona con una puntuación extrema en la primera medición tenderá a presentar su puntuación en la segunda medición más próxima a la media del grupo.

El inevitable fenómeno de regresión a la media tiene una especial relevancia en la crítica de la teoría de la fiabilidad dado que cuestiona la homocedasticidad del modelo.

En términos más sencillos el problema es el siguiente:

1. Toda medición empírica supone una variación *aleatoria* sobre la puntuación verdadera correspondiente.

2. Esta variación aleatoria supone que para una puntuación verdadera V cada una de sus mediciones empíricas correspondientes X tiene *a*

priori tantas probabilidades (p=0'5) de ser mayor como de ser menor que V.

3. Aunque a nivel formal el continuo de capacidad de los sujetos (su ubicación sobre la dimensión que se está midiendo) puede definirse entre menos infinito y más infinito, las puntuaciones de cualquier test real van entre un valor concreto inferior I (mínima puntuación posible, generalmente 0) y un valor concreto superior S (máxima puntuación posible, frecuentemente, para tests con items de respuesta correcta, ésta es el número de items resueltos acertadamente).

4. *Si un sujeto obtiene una puntuación* X *próxima a* S *ó a* I *la próxima vez que lo midamos tiene más probabilidades de obtener una puntuación más cercana a la media del grupo*, más alejada de S ó de I, respectivamente (en esto consiste el fenómeno de la regresión a la media).

5. Para un valor extremo de V, pongamos V igual a S, es imposible que su medición empírica X tenga igual probabilidad de estar por encima que por debajo de V dado que V está ya en el límite de la escala. Esto explica que si se obtiene una puntuación en S o próxima a S la siguiente medición es más probable (de hecho es casi lo único probable debido a cualquier variación aleatoria) que este más lejos del extremo de la escala y por tanto más cerca de la media de las puntuaciones.

6. El modelo clásico trabaja bajo el supuesto de que el modelo es homocedástico, es decir, que presenta igual varianza en cualquier punto de la escala. Sin embargo, la variabilidad de las X entorno a su V, medida por el error típico de medida, no puede ser constante en los extremos de la escala de V. No puede ser constante porque la distribución de las X en torno a su media V ha de estar necesariamente sesgada en los extremos de la escala de V. Este razonamiento cuestiona el modelo explicativo clásico.

La lista anterior de factores que pueden afectar a la segunda medición no es exhaustiva pero contribuye a formar una idea acerca de lo difícil que es, *de hecho,* cumplir las condiciones para una interpretación razonable de la correlación entre dos mediciones como si se tratara de un coeficiente de fiabilidad.

Cada uno de los factores de la lista anterior tiene distinta importancia según el tipo de test de que se trate, el tipo de

sujetos que se esté midiendo y el método de estimación de la fiabilidad que se emplee.

Por ejemplo, los cambios madurativos afectarán especialmente a dimensiones de las aptitudes o la personalidad de naturaleza madurativa en etapas de fuerte desarrollo como niñez o adolescencia y cuando se usa el método test-retest, especialmente en la medida en que el lapso entre mediciones sea amplio. Algunos factores pueden a su vez afectarse mutuamente, potenciarse o compensarse de formas más o menos complejas y anticipables.

Por ejemplo, el cansancio o fatiga debido a la tarea anterior puede actuar junto con efectos atribuibles a la hora del día y del día de la semana (no por la hora o el día en sí, sino por lo que éstos suponen de actividades previas, expectativas, etc. en nuestra cultura). Por ejemplo, en general, no es recomendable administrar pruebas en contextos escolares o industriales el viernes por la tarde o el día antes o después de un puente de fin de semana o de unas vacaciones. Por ejemplo, en general, no deben aplicarse pruebas en contextos educativos después de la hora de educación física. Por ejemplo, debe planificarse la administración de una batería de tests para que las pruebas de aptitudes (inteligencia, factores de inteligencia, atención, etc.) entren en momentos que los sujetos ya están concentrados pero antes de que estén cansados.

Si los sujetos han cambiado o las condiciones han cambiado por uno de los factores anteriores o por una combinación -con posibles interacciones- de los mismos entonces la interpretación de la correlación entre las dos mediciones se vuelve obscura. Esa correlación reflejará indisolublemente todos esos factores además de la fiabilidad del instrumento.

En la mayoría de los casos podría afirmarse razonablemente que los factores que afectan a las condiciones afectarán a las puntuaciones a través de inducir algún cambio, aunque sea momentáneo y pasajero, en los sujetos, en cómo piensan, sienten, interpretan o enfrentan la prueba y, por tanto, en como la contestan. Por esta razón, para abreviar, hablaremos en adelante del ***"cambio en los sujetos"*** como síntesis de cualquier variación debida a los sujetos, a su ambiente social natural o a las condiciones de administración de la prueba. Este cambio en los sujetos se diferencia claramente de la variación aleatoria asociada al instrumento de medida o introducida por el instrumento de medida, a la que llamaremos *variación "debida al instrumento*, y que es de la que trata (o intenta tratar) la fiabilidad.

Con esta aproximación el "cambio en los sujetos" es responsable de cualquier variación en las puntuaciones en

la segunda medición de la que no sea responsable el instrumento.

Como el cambio de los sujetos de una muestra difícilmente es sistemático, simultáneo e igual sujeto a sujeto, en general parece que no puede sostenerse que la segunda medición de cada sujeto es igual a su primera medición más una constante c

$$X_{2i} = X_{1i} + c$$

Aunque, en conjunto, si hay un cambio de signo positivo (en términos algebraicos, no necesariamente psicológicos), es razonable esperar:

$$\bar{X}_2 = \bar{X}_1 + c$$

siendo c tanto más importante cuanto mayor haya sido el cambio de signo positivo.

Por tanto si hay cambio, en general, positivo cabe esperar que:

$$\bar{X}_2 > \bar{X}_1$$

Si el cambio afecta diferencialmente a los sujetos en ambos sentidos (tanto cambio positivo como negativo en promedio, situación improbable para muchas variables) podría no observarse diferencia entre las medias:

$$\bar{X}_2 = \bar{X}_1$$

aunque, sin embargo cabría esperar razonablemente -aunque tampoco con seguridad- que

$$s_{X_1} \neq s_{X_2}$$

Una situación con cambio tal que medias y desviaciones permanecieran constantes es teóricamente posible pero prácticamente inverosímil.

Por ejemplo, si hay un efecto de la fatiga sobre los sujetos, las puntuaciones de la forma B, o segunda medición, serán, en general, más bajas. Si las puntuaciones de la forma B son más bajas, entonces la media empírica de la forma B será menor que la de la forma A. Es decir, la media de las empíricas de la forma B será igual a la media de las empíricas de la forma A menos una constante c:

$$\bar{X}_A - c = \bar{X}_B$$

o, dicho de otro modo:

$$\bar{X}_A = \bar{X}_B + c$$

Según la teoría, la media de las empíricas ha de ser la media de las verdaderas, por tanto:

$$\bar{V}_A > \bar{V}_B$$

de lo que puede concluirse claramente que las formas A y B no son paralelas. Este efecto puede verificarse empíricamente, basta comprobar si las medias de las empíricas de ambas formas son iguales.

Al contrario de lo que podría pensarse intuitivamente, aunque se produzca cambio este no es necesariamente detectado por el coeficiente de fiabilidad. Como el coeficiente de fiabilidad así estimado se calcula como la correlación entre las puntuaciones observables de los sujetos en las dos formas, este podría no detectar un cambio lineal de las puntuaciones debido a fatiga u otras razones semejantes.

Por otra parte, debido a que la medición real tiene límites en su escala en los valores I y S, cualquier desplazamiento de la distribución que la aproxime a alguna cola tiende a truncar o recortar la distribución acumulando casos en esa cola, lo que afecta a la forma de la distribución y puede afectar claramente a su dispersión. Esto cuestionaría la igualdad de varianzas que exige la paralelidad. De nuevo esto puede verificarse empíricamente.

5. El diseño de mediciones paralelas

La lógica de los métodos de estimación de la fiabilidad es, supuestamente, experimental. De hecho, por ejemplo Gulliksen (1950), los denomina literalmente así “Métodos Experimentales para la obtención de la fiabilidad”. Por ese motivo es claro que pueden analizarse a la luz del diseño de investigación que implican, al margen de las consideraciones prácticas y de las consideraciones estadísticas.

El razonamiento de estos métodos, desde el punto de vista del diseño es como sigue. Si efectuamos dos mediciones de una muestra de sujetos sin que los sujetos cambien, sin que las condiciones cambien, y sin que el instrumento de medida cambie, entonces debemos esperar los mismos resultados salvo por variación debida al azar. Si el instrumento no introduce error (variación aleatoria) en la medición, entonces los resultados de dos mediciones en condiciones controladas deberá ser el mismo.

¿Que mecanismos experimentales de evitación de efectos espurios se utilizan?

> En primer lugar, se pretende un *control de condiciones experimentales,* manteniendo constantes e irrelevantes todas las variables (incluido el instrumento de medida). Este control de condiciones es siempre relativo dado que los tests, por lo general,

no se administran en contextos de laboratorio, sino en contextos sociales reales (clínica, colegio, empresa, universidad).

En segundo lugar, puede suponerse razonablemente que no hay efectos espurios importados con la selección de los sujetos si estos son una *muestra aleatoria* de la población de interés.

Sin embargo, a diferencia de un diseño experimental *no hay grupo control y tampoco hay manipulación de la variable independiente,* no hay "tratamiento" alguno.

La única diferencia entre la primera medición y la segunda medición ha de ser justamente esa, que una es la primera medición y la otra la segunda, y ambas han de ser dos mediciones completamente independientes entre sí.

En la notación de diseños experimentales de Campbell y Stanley (1978) el diseño de investigación que se utiliza para estimar la fiabilidad de un test mediante mediciones paralelas (sea formas paralelas, test-retest o mitades) tiene la forma:

R	O_1	O_2

Es decir, se obtiene una muestra aleatoria (R) de sujetos de la población de interés y, bajo condiciones controladas se la mide (O_1) y se la vuelve a medir (O_2).

En términos de diseño clásico es obvio que este diseño carece de validez interna y podría calificarse de pseudo o pre-experimental.

Teóricamente, dado que no se ha introducido ningún factor de cambio, no habría que esperar ningún cambio en las puntuaciones entre O_1 y O_2 pero, *de hecho,* esos cambios aparecen y, cuando aparecen con este diseño no pueden atribuirse claramente a nada.

La dudosa lógica con la que habitualmente se trata este diseño es: nada ha cambiado (ni condiciones, ni sujetos, ni el instrumento de medida), por tanto no deberían haber cambios en las puntuaciones, si los hay es que el instrumento no es fiable, es decir, es que el instrumento *produce* la variación aleatoria. Este último razonamiento no es lógicamente correcto.

Si el instrumento no cambia en absoluto, y de eso estamos seguros, en buena lógica experimental, el instrumento no puede ser un factor que explique la variación de las puntuaciones.

Puede argumentarse que efectivamente no es el instrumento el que introduce las variaciones aleatorias

que reducen la fiabilidad de la medición. Pero, si no se pueden atribuir al instrumento sino a un acto particular de medición ¿cómo hablar de la fiabilidad *del* test?

Si se piensa en un ejemplo concreto podrá verse con mayor claridad el fallo lógico del modo de razonar asociado al diseño de mediciones paralelas. Supongamos, para ponerle las cosas fáciles al modelo clásico, que queremos medir la capacidad para hacer "divisiones con cuatro cifras en el dividendo y dos en el divisor", una prueba adecuada para niños de cierto grado educativo. Construimos una prueba de 30 ítems (30 divisiones). Supongamos una muestra suficiente de N sujetos extraída aleatoriamente de la población de interés. La misma psicóloga, experta en el uso de tests, administra la prueba a esta muestra dos veces, con un lapso de cuatro días. Los sujetos no reciben entrenamiento en divisiones ni aspectos relacionados en ese periodo, ni ningún otro tratamiento especial. El lapso de cuatro días parece razonablemente adecuado para evitar que efectos de maduración, memoria, o aprendizaje afecten de un modo importante, según pruebas anteriores. Supongamos que todas las condiciones ambientales se han mantenido cuidadosamente constantes.

Supongamos por un momento que aparecen diferencias significativas entre las medias empíricas de las dos mediciones. ¿A que atribuir estas diferencias? Desde un razonamiento experimental si estuviéramos seguros de que no ha variado nada entonces a nada podemos atribuirlas. Por ejemplo, si estamos seguros que no ha variado el administrador del test no podríamos atribuir a este las diferencias. Si no han variado las instrucciones, ni el tiempo, ni la hora del día, ni el local, ni la luz ... no podríamos atribuir los cambios a ninguno de esos factores. Del mismo modo, *si estamos seguros que el test no ha cambiado ¿cómo podríamos atribuir los cambios al test?* Este razonamiento absurdo es el que parece sostener la teoría de la fiabilidad.

Aun en las condiciones idóneas de este ejemplo, con un diseño como éste no tenemos ninguna garantía de que otras variables no hayan influido para producir las variaciones sistemáticas o aleatorias que aparezcan en las puntuaciones de la segunda medición. Por ejemplo, no hay garantía de que la primera aplicación no produzca efectos sobre la segunda, vía aprendizaje, práctica, memoria de resultados, etc. Tampoco hay garantía de que los sujetos no hayan madurado de algún modo que afecte sus resultados, o de que no hayan cambiado realmente, por la razón propia o ambiental que sea, en la variable de interés.

Estos razonamientos llevan a una conclusión principal: *El diseño de mediciones repetidas de la teoría clásica de tests es incapaz de distinguir entre cambios debidos a los sujetos y cambios debidos al instrumento. Por esta razón este diseño de investigación debe ser abandonado dado que no es adecuado para el propósito de establecer la fiabilidad.*

Puede argumentarse que la teoría de la fiabilidad no sostiene que el instrumento *produce* la variación aleatoria que reduce la correlación entre mediciones por debajo de 1. Pero, si no la produce el instrumento ¿cómo decir que ésta es la fiabilidad *del* instrumento de medida? Ante este dilema la teoría de la fiabilidad no tiene más que dos opciones. O admite que el instrumento es el que *produce* la variación, lo que permite hablar de la fiabilidad *del* instrumento pero supone admitir un razonamiento experimental absurdo. O bien admite que no se conoce el origen de la variación, pudiendo atribuirse al instrumento, a los sujetos o a las condiciones o a cualquier combinación e interacción de variables de todos estos tipos, lo que supone un razonamiento experimentalmente más realista pero desautoriza la idea de que la correlación entre las dos mediciones es la fiabilidad del instrumento. Cualquiera de las dos líneas argumentales cuestiona el modo en que se realiza la estimación del llamado coeficiente de fiabilidad y desautoriza el uso habitual del mismo preconizado alegremente de manual en manual.

6. Paralelidad y cuantía del coeficiente de fiabilidad

Si existen diferencias significativas entre medias, o entre varianzas de las puntuaciones empíricas obtenidas, las mediciones ya no son paralelas y el resultado de la correlación entre las puntuaciones empíricas no es por tanto el coeficiente de fiabilidad. La teoría sólo garantiza que *si las mediciones son paralelas* su correlación es el coeficiente de fiabilidad.

El coeficiente de fiabilidad no mide por tanto en qué grado son paralelas las pruebas (y por tanto, en su caso, puede resultar confuso denominar a la correlación entre formas paralelas coeficiente de equivalencia).

> *Si las mediciones son paralelas su correlación (sea cual sea, teóricamente sea alta o sea baja) es el coeficiente de fiabilidad.*
>
> Si las mediciones no son paralelas su correlación (aunque sea alta) no es el coeficiente de fiabilidad. Paradójicamente dos formas supuestamente paralelas de un test podrían correlacionar 1 y no ser paralelas: Basta que una (Y) sea función lineal perfecta (Y=a+bX) de la otra (X) con los parámetros "a" y "b" de

la ecuación de regresión que se quiera, con la única restricción de que "a" y "b" no permitan Y=X, siendo b positivo.

Esta última conclusión resultará sorprendente a muchos (aunque no es más que una lectura del modelo de mediciones paralelas) porque en la práctica se procede *como si* las pruebas fueran paralelas independientemente de que lo sean o no.

Este uso generalizado supone que muchos resultados que se presentan como el coeficiente de fiabilidad de tests en realidad no lo son, o, en el mejor de los casos no se sabe si lo son debido a que no se ha contrastado si se dispone de mediciones paralelas.

Si pudiéramos establecer formalmente que los supuestos de paralelidad se cumplen escrupulosamente ello no supondría necesariamente que el coeficiente de fiabilidad deba ser alto. En teoría la variación aleatoria podría disminuir sustancialmente esa correlación.

Que dos mediciones (tests, partes, test-retest) sean o no sean paralelas no depende de la cuantía de la correlación entre ellas. Aunque si tenemos dos formas paralelas sería razonable esperar, en la práctica, correlaciones positivas muy elevadas.

7. Ausencia de contrastación de la paralelidad y cambio lineal

Este modo de proceder generalizado (calcular el coeficiente de correlación y considerarlo estimación de la fiabilidad sin contrastar la paralelidad) tiene más consecuencias que no deben pasar inadvertidas.

Supongamos que por algún motivo los sujetos de una muestra cambian en la variable de interés por razones madurativas, ambientales o del tipo que sea entre la primera y la segunda medición.

Supongamos que ese cambio puede describirse razonablemente con la función:

$$X_B = c_1 + c_2 X_A$$

Es decir, la puntuación en la forma B es igual a la puntuación en la forma A multiplicada por una constante c_2, más una constante c_1. Este es el tipo función más sencilla y usual cuando se trata de pronosticar una variable por otra.

Es decir que las puntuaciones en la segunda medición son función lineal de las puntuaciones en la primera medición.

El coeficiente de correlación de Pearson es invariante ante transformaciones lineales de una o ambas de las variables consideradas (véase por ejemplo Amón (1979) para una demostración).

El coeficiente de correlación de Pearson por tanto *no puede detectar un cambio lineal.*

Si el investigador calcula la correlación entre las puntuaciones A y B obtendrá una correlación de 1. "¡Perfecto!" diría un investigador convencional, "el test es perfectamente fiable".

En realidad se ha ignorado el cambio. El test podría estar arrojando ahora, por ejemplo, puntuaciones que respondieran a la fórmula:

$$X_B = 1'5 + 3\,X_A$$

Sería difícil sostener que un test así, supuesto que pudiera descartarse cualquier otro origen del cambio, es "fiable". ¿Cuál será el cambio para una tercera medición? Desde luego el test no estaría produciendo las mismas "puntuaciones verdaderas", aunque las nuevas verdaderas serían también función lineal de las antiguas.

Sin embargo, si la relación lineal existe, pero no es perfecta

$$X'_B = c_1 + c_2 X_A$$

podemos obtener sustanciosos supuestos coeficientes de fiabilidad que den la impresión de un test fiable cuando en realidad no lo es.

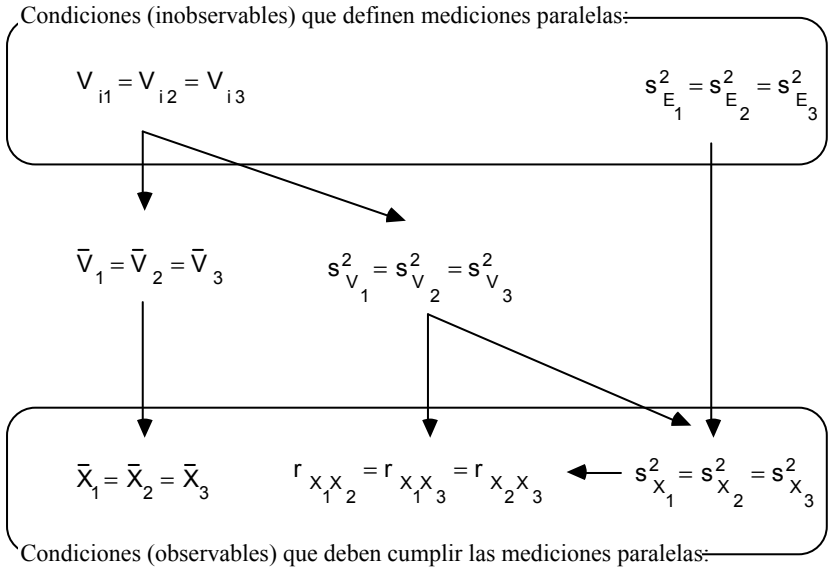
Un coeficiente de correlación positivo y cercano a 1 entre dos mediciones garantiza que una medición es aproximadamente una función lineal de la otra, pero no garantiza ni que tengamos dos mediciones paralelas ni que el instrumento sea fiable. Pueden haber aparecido cambios lineales entre una y otra medición, debidos al test, a los sujetos o a diversas condiciones, y estos cambios lineales no serán detectados por la correlación.

8. La definición de las mediciones paralelas no puede verificarse empíricamente

Tradicionalmente la correlación entre dos tests paralelos (dos tests o formas distintas de un test que cumplen las condiciones de paralelidad) se presenta como el punto de enganche entre la Teoría Clásica de Tests y la medición real. Obsérvese que sin ese punto de contacto toda la teoría anterior carece de sentido para el trabajo práctico porque no puede operarse con ella.

En el cuadro siguiente pueden verse las deducciones básicas sobre el concepto de mediciones paralelas. Cada flecha del cuadro representa una deducción en función de fórmulas del modelo. Hemos explicado con anterioridad todas esas deducciones, excepto la que lleva a afirmar que las correlaciones (coeficientes de fiabilidad) entre tests paralelos han de ser todas iguales entre si, pero esta deducción es inmediata de la definición de correlación entre formas paralelas como cociente entre la varianza verdadera y la empírica, que, por deducciones anteriores han de ser iguales.

DEDUCCIONES BASICAS SOBRE MEDICIONES PARALELAS



El problema es que las dos condiciones de paralelidad:

$$V_{i1} = V_{i2} = V_{i3}$$

$$s^2_{E_1} = s^2_{E_2} = s^2_{E_3}$$

se refieren ambas a inobservables que sólo pueden ser estimados una vez que disponemos del coeficiente fiabilidad, pero necesitamos establecer que dos mediciones son paralelas para poder considerar que su correlación es el coeficiente de fiabilidad.

Por este motivo tiene interés considerar si se puede deducir la paralelidad de las medidas a partir de condiciones observables.

Si se cumplen *las tres* condiciones observables:

$$\bar{X}_1 = \bar{X}_2 = \bar{X}_3$$

$$s^2_{X_1} = s^2_{X_2} = s^2_{X_3}$$

$$r_{X_1X_2} = r_{X_1X_3} = r_{X_2X_3}$$

entonces todos los pasos del esquema son reversibles, excepto el que lleva a afirmar que:

$$V_{i1} = V_{i2} = V_{i3}$$

Es decir, si tenemos tres o más mediciones que tienen iguales medias, varianzas y correlaciones entre sí, supuesto que el modelo definido por

$$X_i = V_i + E_i$$

$$\bar{E} = 0$$

sea cierto, entonces, podemos llegar a deducir que en esas mediciones las medias y las varianzas verdaderas son iguales:

$$\bar{V}_1 = \bar{V}_2 = \bar{V}_3$$

$$s^2_{V_1} = s^2_{V_2} = s^2_{V_3}$$

Pero no podemos llegar a afirmar que:

$$V_{i1} = V_{i2} = V_{i3}$$

Aunque si dos variables tienen todas sus puntuaciones iguales caso a caso ello implica que tienen igual media e igual varianza, lo inverso no es necesariamente cierto. Si dos variables tienen igual media y varianza ello no

implica que necesariamente tengan las mismas puntuaciones caso a caso.

En mi opinión este es un punto esencial que los manuales de teoría de tests soslayan.

> Primero, la definición de mediciones paralelas es esencial porque se define la fiabilidad como correlación entre mediciones paralelas. Segundo, por tanto, es esencial dilucidar cuando dos mediciones son paralelas. Tercero, la definición de mediciones paralelas no puede verificarse (aunque si falsarse) empíricamente.

La conclusión principal es que no podemos saber si estamos estimando un coeficiente de fiabilidad cuando correlacionamos dos mediciones nominalmente paralelas (aunque si se incumplen las condiciones observables sí podemos saber que no son mediciones paralelas y por tanto que no estamos estimando un coeficiente de fiabilidad).

El error lógico de identificar el contraste a nivel de puntuaciones empíricas con la satisfacción de los criterios de paralelidad

El problema es que no podemos saber si dos mediciones tienen puntuaciones verdaderas iguales y varianzas de error iguales porque justamente las puntuaciones verdaderas y las varianzas de error no son observables.

Lo máximo que podemos hacer es comprobar algunas deducciones del modelo, como que las medias de las puntuaciones empíricas sean iguales entre sí y que las varianzas de las puntuaciones empíricas sean iguales entre sí. Esto, es condición necesaria pero no suficiente para que las mediciones sean paralelas.

En mi opinión hay un error lógico en el tratamiento habitual de este tema en algunos manuales al uso, nacionales e importados, que consiste en identificar el contraste de dos condiciones observables de paralelidad (ausencia de diferencias estadísticamente significativas entre las medias, por un lado, y las varianzas, por otro, de las puntuaciones empíricas de dos mediciones supuestamente paralelas.) con las condiciones inobservables de paralelidad: puntuaciones verdaderas iguales y varianzas de error iguales

Las medias de dos mediciones podrían ser iguales entre sí y las varianzas iguales entre sí y no ser cierto que las puntuaciones verdaderas fueran iguales entre si sujeto a

sujeto ni que la varianza de error fuera igual entre ambas mediciones. En realidad de ahí ni siquiera se desprende que la puntuación empírica pueda descomponerse en puntuación verdadera y error de medida. Otros muchos modelos serían compatibles con este resultado. Para deducir desde lo observable hasta lo inobservable hay que seguir suponiendo que el modelo es cierto.

De las dos condiciones de paralelidad que establece el modelo (puntuaciones verdaderas iguales y varianzas de error iguales) se desprende que las medias y las varianzas de las puntuaciones empíricas de dos mediciones paralelas han de ser iguales. Si no lo son no estamos ante formas paralelas, pero si lo son esto no garantiza que estemos ante mediciones paralelas.

De la primera condición de paralelidad (puntuaciones verdaderas iguales) se desprende que las medias de las puntuaciones verdaderas han de ser iguales y de ahí que las medias de las empíricas han de ser iguales. Si no lo son no estamos ante mediciones paralelas, pero si las medias empíricas son iguales esto no garantiza que estemos ante mediciones paralelas.

Aunque pudiéramos comprobar -que no podemos directamente- la ausencia de diferencias significativas entre las medias y las varianzas de las puntuaciones verdaderas, esto no garantizaría que las puntuaciones verdaderas fueran iguales sujeto a sujeto, como pretende el modelo.

Papel del contraste estadístico de las condiciones observables

En mi opinión el contraste estadístico de las condiciones observables (igualdad de medias, de varianzas, de correlaciones) no sirve para confirmar la paralelidad de las mediciones, pero si para descartar que las mediciones sean paralelas.

Dado que la igualdad de medias, de varianzas y de correlaciones es condición necesaria para que las mediciones sean paralelas, si se incumple alguno de estos criterios se sabe que las mediciones no son paralelas.

El modelo exige la igualdad, pero debido a que por razones de error muestral incluso mediciones extraídas al azar de una población de mediciones paralelas presentarían diferencias entre sí, resulta adecuado enfocar la cuestión como un contraste estadístico. Determinar si las mediciones son iguales más allá del grado de variabilidad atribuible al error muestral.

Número de mediciones supuestamente paralelas necesarias

Si sólo comprobamos si las medias empíricas son iguales y si las varianzas empíricas son iguales, entonces, supuesto que el modelo es cierto, ni siquiera podemos deducir si las

varianzas verdaderas son iguales, y mucho menos, claro, que las puntuaciones verdaderas sean iguales caso a caso.

La consecuencia de esto es que para llegar a comprobar, por lo menos, que las medias de las puntuaciones verdaderas son iguales y que las varianzas de las puntuaciones verdaderas son iguales hace falta disponer de al menos tres formas paralelas para verificar que sus medias, sus varianzas y sus correlaciones empíricas son iguales.

Plantear un contraste solo para comprobar si las medias y las varianzas son iguales permitirá descartar como paralelas aquellas mediciones en que la igualdad no se cumpla, pero si se cumple, ni siquiera podrá afirmarse que las varianzas de las puntuaciones verdaderas de las mediciones sean iguales.

El razonamiento tautológico clásico

En este punto relativo a establecer si dos mediciones son o no paralelas, la mayoría de manuales efectúan un razonamiento tautológico implícito en la Teoría Clásica de Tests. Este razonamiento comienza suponiendo que dos mediciones son paralelas. Como son paralelas entonces su correlación es el coeficiente de fiabilidad. Una vez obtenido el coeficiente de fiabilidad puede calcularse el error típico de medida, la varianza de error, la varianza verdadera, etc.

La contradicción reside en que la correlación entre dos mediciones es igual al coeficiente de fiabilidad *si* las mediciones son paralelas (es decir, iguales puntuaciones verdaderas e iguales varianzas de error). Si no son paralelas su correlación no es el coeficiente de fiabilidad.

Esta cuestión, que en mi opinión es central, es habitualmente soslayada. En algún punto comienza a procederse *como si* la definición de tests paralelos hubiera sido satisfecha y por tanto se pudiese calcular realmente el coeficiente de fiabilidad. A lo sumo se admite que no se puede asegurar que dos mediciones sean paralelas (p.e. Crocker y Algina, 1986), pero inmediatamente se procede como si lo fueran hablando de estimaciones muestrales del coeficiente de fiabilidad.

Explicación adicional:

Si al efectuar O_2 tenemos que:

$$\bar{X}_2 = \bar{X}_1$$

y que:

$$s_{X_1} = s_{X_2}$$

podría parecer razonable pensar que las mediciones pueden ser paralelas. Pero ello no es necesariamente así.

Las condiciones de paralelidad son dos:

$$V_{i1} = V_{i2}$$

$$s^2_{E_1} = s^2_{E_2}$$

Si las mediciones cumplen las dos condiciones de paralelidad entonces cumplirán también qué:

$$\bar{X}_2 = \bar{X}_1$$

$$s_{X_1} = s_{X_2}$$

que son dos condiciones observables.

Pero si las mediciones cumplen que:

$$\bar{X}_2 = \bar{X}_1$$

$$s_{X_1} = s_{X_2}$$

entonces no necesariamente cumplen:

$$V_{i1} = V_{i2}$$

$$s^2_{E_1} = s^2_{E_2}$$

Y por tanto no necesariamente son paralelas.

Un ejemplo puede ilustrar intuitivamente esta cuestión. Veamos los siguientes datos hipotéticos:

	0_1	0_2
S_1	9	1
S_2	9	9
S_3	9	1
S_4	9	9
S_5	5	5
S_6	5	5
S_7	1	1
S_8	1	9
S_9	1	1
S_{10}	1	9

El lector puede comprobar fácilmente (se aprecia a simple vista porque hay los mismos números en las dos columnas) que ambas mediciones tienen la misma media (5'0) y la misma varianza (3'57). Sin embargo, es difícil sostener que las puntuaciones de cada sujeto expresan la misma puntuación verdadera. La correlación es 0.

Puede argumentarse que tales mediciones podrían considerarse mediciones paralelas con fiabilidad 0, pero, aparte de que esto no tiene mucho sentido, bastaría hacer modificaciones de grado en los números anteriores para, manteniendo medias y varianzas iguales obtener coeficientes de correlación distintos de 0 sin que tengamos criterio para afirmar que se trate de mediciones paralelas.

También pueden construirse contraejemplos con 3 variables, lo que permite sostener simultáneamente las correlaciones iguales, medias iguales y varianzas iguales sin que resulte razonable suponer que provienen de puntuaciones verdaderas iguales.

Criterio de contenido psicológico

Además, cuando las supuestas mediciones paralelas están formadas por items distintos (todos los casos de mediciones paralelas excepto test-retest) habría que garantizar, razonablemente, la paralelidad de contenido psicológico entre los items. Esto significa que los items han de ser muestra o signo de los mismos contenidos psicológicos.

El criterio de igual contenido psicológico es necesario incluso aunque se cumplan todas las garantías de paralelidad estadísticas posibles.

Obsérvese que, al menos teóricamente, podríamos elaborar, paradójicamente, medidas perfectamente paralelas entre sí según criterio estadístico, pero referidas a variables psicológicas completamente distintas (lo cual, obviamente no tiene sentido y contradice el sentido del concepto "fiabilidad" de *una* medida).

Orientación del contraste estadístico

En mi opinión, no puede plantearse el contraste estadístico entre medias, varianzas o correlaciones de supuestas mediciones paralelas como quien se plantea, por ejemplo, si varones y mujeres difieren en su inteligencia verbal.

La ausencia de diferencias significativas efectuando un contraste tradicional de medias ni siquiera garantiza que no haya una diferencia importante y desde luego no garantiza la igualdad que exige el modelo.

No poder rechazar la hipótesis nula de igualdad de medias, por ejemplo, significa que la diferencia observada entre medias podría deberse al azar debido a variación muestral. Significa que no hay evidencia para rechazar la idea de que ambos grupos de puntuaciones provengan de una misma población con una misma media. Sin embargo, no poder rechazar la hipótesis de igualdad (de medias, de varianzas, de correlaciones) no equivale a probar la igualdad (de medias, de varianzas o de correlaciones).

Aunque, por ejemplo, encontráramos una ausencia de diferencias significativas con un 99% de confianza, esto no significaría, que pudiéramos considerar como iguales ambas medias. Solo significaría que una diferencia entre medias como la hallada pertenece al 99% de diferencias más común halladas cuando la hipótesis nula es cierta, es decir, que no es una diferencia tan grande que aparezca tan solo un 1% de las veces o menos por azar cuando la hipótesis nula es cierta (es decir, cuando se sabe que no hay diferencia real). Por ejemplo una diferencia no significativa al 99% podría haber sido considerada significativa al 95%.

¿Cómo vamos a admitir que dos mediciones son iguales porque su diferencia no sea tan grande que aparezca por azar un 1% ó un 5% de las veces? Así bien poco exigentes seríamos para admitir que dos medidas fueran paralelas. Si se hace un contraste con enfoque tradicional y se aplica un nivel de confianza del 99% prácticamente todas las mediciones resultarán paralelas. Pero si se es “más exigente” en la linea tradicional y se aplica un nivel de confianza compulsivo del 1 por 10.000 hay garantía de que no encontraremos mediciones que no sean paralelas.

Dado que lo que trata de probar el investigador es la igualdad y no la diferencia entre medias, garantizar un resultado aceptable no atribuible al azar implicaría plantear la zona de rechazo de la hipótesis al revés de lo usual. Habría que determinar una zona de aceptación de la hipótesis de igualdad no debida al azar del 5% en torno a la ausencia de diferencias. De ese modo podría afirmarse, si no se rechaza la hipótesis de igualdad, que mediciones tan semejantes solo aparecen por azar el 5% de las veces o menos. Admitiendo que sólo se conoce la distribución del estadístico bajo hipótesis nula podría afirmarse que mediciones tan semejantes solo aparecen por azar el 5% de las veces o menos cuando las muestras provienen de una población con la misma media. Admitiendo que sólo se

conoce la distribución del estadístico bajo hipótesis nula nada puede decirse de la probabilidad de aparición de cierta diferencia cuando las muestras proceden de poblaciones distintas con parámetros desconocidos.

En términos de tipos de error de contraste de hipótesis, dado que lo que deseamos probar es que los dos estadísticos son iguales (sean las medias, las varianzas o las correlaciones) el error que debe preocuparnos esencialmente no es el error Tipo I (probabilidad de rechazar como falsa una hipótesis nula que es cierta, es decir, la probabilidad de que diéramos por diferentes medias o varianzas que realmente son iguales) sino el error Tipo II (probabilidad de aceptar como cierta una hipótesis nula que realmente es falsa, es decir, la probabilidad de dar por iguales estadísticos que en realidad no lo son). De ese modo quizás podríamos admitir que los dos estadísticos provienen de un mismo parámetro porque una igualdad tan estrecha sólo se da por azar el 5% de las veces o menos cuando realmente ambos provienen de un mismo valor del parámetro.

Si las medias, varianzas y correlaciones no difieren significativamente es posible proceder "como si" estos estadísticos fueran iguales en la población, puesto que no han podido rechazarse las hipótesis nulas cuya veracidad es condición necesaria pero no suficiente para la igualdad. Pero hay que ser conscientes de que no se ha probado la igualdad y por tanto el ajuste del modelo a los datos no ha sido verificado.

> En conjunto mis opiniones y razonamientos en todas estas cuestiones de la fiabilidad son críticos y al menos parcialmente contrarios a la doctrina tradicional, pero en este punto del contraste estadístico de la paralelidad además pueden ser especialmente polémicos o simplemente erróneos. El lector puede encontrar un enfoque de la cuestión exactamente al revés en otros manuales. Sin embargo, algunos comentarios de Gulliksen (1950) acerca de orientar el contraste hacia el contraste no de "diferencias significativas", sino de "diferencias despreciables" parecen respaldar mi punto de vista.

9. El criterio estadístico de Wilks para tests paralelos

Gulliksen (1950) dedicó el capítulo 14 de su manual, justamente a discutir "un criterio estadístico para tests paralelos". Para ello utilizó como *definición operativa de tests paralelos* aquellos que tienen iguales medias, varianzas y correlaciones entre ellos. Gulliksen (1950)

expone el tema siguiendo a Wilks (1946) que había desarrollado estadísticos para contrastar la igualdad de medias, varianzas y correlaciones entre supuestos tests paralelos, y siguiendo a Votaw (1947, 1948) que había desarrollado estadísticos para contrastar que un conjunto de tests paralelos han de presentar la misma correlación (validez) con un criterio externo al test (una variable de referencia con la que por hipótesis se espera determinada correlación, muchas veces positiva y alta).

Ya he discutido el significado y limitaciones que, en mi opinión, puede tener este tipo de contraste, ahora vamos a presentar el estadístico de Wilks para contrastar paralelidad. El método requiere 3 o más formas paralelas. Dado que habitualmente construir tres formas paralelas ya está más allá de lo usual, nos centraremos sobre el caso mínimo. Si no hubiera al menos 3 formas paralelas no se podría contrastar la igualdad entre correlaciones entre formas, y, como hemos visto, si no se contrasta esto ni si quiera se puede garantizar que las varianzas de error sean iguales.

Si tenemos tres mediciones nominalmente paralelas, el estadístico L_{mvc} pone a prueba simultáneamente la hipótesis de que las tres medias empíricas son iguales entre si, las tres varianzas empíricas son iguales entre si y las tres covarianzas empíricas son iguales entre si.

$$L_{mvc} = \frac{s_1^2 s_2^2 s_3^2 (1 + 2r_{12} r_{13} r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2)}{s^2 (1 + 2r)(s^2 - s^2 r + v)^2}$$

En la fórmula,

s^2 representa la varianza media de las tres mediciones (es decir la suma de las tres varianzas dividida por 3).

r representa la correlación media entre las tres mediciones. Esta correlación media se puede obtener calculando la covarianza media y dividiéndola por la varianza media.

v es la varianza de las tres medias (para lo que hace falta calcular la media de las 3 medias, restarla de cada media, (sumar las diferencias elevadas al cuadrado) y dividir por 3).

Los demás términos son varianzas de cada forma o correlaciones entre formas.

Según Wilks, para N>100 el estadístico $-N \cdot \log L_{mvc}$ distribuye aproximadamente como ji-cuadrado con 6 grados de libertad (para el caso de 3 tests).

Según Gulliksen, L_{mvc} varía entre 0 y 1. Si las medias son iguales, las varianzas son iguales y las correlaciones son iguales, el estadístico vale 1. Según L_{mvc} se aproxima a 1, $-N \cdot \log L_{mvc}$ se aproxima a 0.

Para el caso concreto de 3 tests, que es el que estamos considerando por razones prácticas si el estadístico $-N \cdot \log_{10} L_{mvc}$ calculado sobre una muestra de al menos 100 casos, es menor que el valor crítico 5'47 podemos afirmar que los tests son paralelos a un nivel $\alpha = 0'05$.

Para el caso de 3 tests, si el estadístico $-N \cdot \log_{10} L_{mvc}$ calculado sobre una muestra de al menos 100 casos, es menor que el valor crítico 7'3 podemos afirmar que los tests son paralelos a un nivel $\alpha = 0'01$.

Si este estadístico fuera mayor que 7'3 hay tan solo una probabilidad sobre 100 o menos de que las tres formas provengan de una población donde las medias sean iguales, las varianzas iguales y las correlaciones iguales.

A partir de este estadístico Wilks desarrolló otro que evalúa únicamente la igualdad de varianzas y de correlaciones, y un tercero, que evalúa exclusivamente la igualdad de las medias, con el propósito de que sirvan a un diagnóstico más fino cuando L_{mvc} muestra que las formas no son paralelas. No las detallamos aquí porque es suficiente el incumplimiento del criterio establecido por L_{mvc} para que los tests no sean paralelos

A pesar de que L_{mvc} es, al parecer, el único test estadístico que se ha desarrollado específicamente para someter a contraste la paralelidad de supuestas formas paralelas, por alguna razón que desconozco no ha hecho fortuna más allá de la obra de Gulliksen y los manuales posteriores han tendido a ignorarlo. Quizás exista evidencia de que el estadístico no opera como debiera. En este sentido algunos ensayos empíricos no sistemáticos con muestras pequeñas nos han mostrado que el estadístico podría fallar en algunos casos en identificar formas paralelas.

El estadístico L_{mvc} de Wilks pone a prueba si los tres tests o formas satisfacen la definición operativa de tests paralelos: iguales medias empíricas, iguales varianzas empíricas e iguales correlaciones entre ellos. Esta definición operativa de tests paralelos si puede ser plenamente contrastada pero, como hemos visto es condición necesaria, pero no suficiente para cumplir el criterio de paralelidad de la teoría clásica (iguales puntuaciones verdaderas sujeto a sujeto e igual varianza de error)

En términos teóricos este desfase entre la definición operativa y la definición teórica de paralelidad tiene importancia para juzgar la teoría clásica de tests y su capacidad de ser contrastada y útil. En términos prácticos, la satisfacción de la definición operativa de paralelidad y la satisfacción del criterio de contenido psicológico puede ser

más que suficiente para poder trabajar razonablemente con dos tests o formas como si fueran formas paralelas.

> Nuestros razonamientos parecen acordes en esta cuestión con los de Lord y Novick (1968; pag. 59) para quienes el proceso de determinar si dos o más tests son suficientemente paralelos requiere la verificación de iguales medias empíricas, iguales varianzas empíricas, iguales correlaciones entre formas, e iguales correlaciones de cada forma con un criterio externo. Pero, en sus propios términos: "La validez de esas ecuaciones para la población en cuestión es una condición de paralelismo necesaria pero no suficiente", aunque "es suficiente para establecer la aplicabilidad de los resultados estándar a esa población."

10. Mediciones paralelas, cambio y fiabilidad

En síntesis el modelo de mediciones paralelas que sustenta el cálculo de la fiabilidad presenta la siguiente lógica y particularidades principales.

Pasos:

Se obtiene una muestra aleatoria de sujetos y se mide dos o más veces en condiciones constantes: O_1 O_2 y O_3.

Se evalúa la relación entre cada dos observaciones con un coeficiente de correlación de Pearson.

Interpretación de la correlación entre mediciones:

Si no hay ninguna variación entre dos mediciones el coeficiente de correlación entre ellas valdrá 1.

Cualquier variación no lineal entre dos mediciones producirá un decremento de ese coeficiente de correlación.

Cualquier variación lineal entre dos mediciones no afectará a ese coeficiente de correlación.

Si dos mediciones son mediciones paralelas, entonces deben cumplir dos condiciones directamente contrastables: Medias iguales y varianzas iguales. Si

las cumplen las mediciones pueden ser paralelas; si incumplen una las mediciones no son paralelas. Si se cumple la igualdad de medias y la de varianzas esta garantiza razonablemente la ausencia de cambio lineal entre mediciones, en el caso común de que la correlación entre ambas mediciones sea considerable y de signo positivo.[1]

Si tres o más mediciones son mediciones paralelas, entonces deben cumplir tres condiciones directamente contrastables: Medias iguales, varianzas iguales y correlaciones iguales entre mediciones. Si cumplen las tres condiciones las mediciones pueden ser paralelas; si incumplen una las mediciones no son paralelas. Si se cumplen las tres condiciones la igualdad de medias y la de varianzas garantiza que no ha aparecido cambio lineal entre ellas.

Si se cumplen las condiciones observables de paralelidad entonces la correlación entre mediciones podría ser una estimación de la fiabilidad del instrumento.

En términos prácticos, si se cumplen estas condiciones podría considerarse *que posiblemente* las mediciones sean paralelas y que los factores que pueden afectar la segunda medición no la han afectado. (Aunque este razonamiento, como hemos mostrado, no es riguroso y puede estar completamente equivocado.)

[1] Como ya hemos señalado, la igualdad de medias y la de varianzas, por sí, no garantizan un comportamiento de los datos consistente con una declaración de fiabilidad (ver ejemplo de la página 82). De hecho, hipotéticamente podría darse también el caso de dos variables O_1 y O_2, con medias empíricas iguales y varianzas empíricas iguales entre si, entre las que hubiese un cambio lineal negativo perfecto. En efecto, si, por simplicidad, denominamos X a O_1 e Y a O_2, si $Y= (X_{máximo} + X_{mínimo}) - X$, ecuación de regresión en la que $a=(X_{máximo} + X_{mínimo})$ y $b=-1$, tal que X e Y contengan los mismos datos, entonces necesariamente las medias de X e Y son iguales, y las varianzas de X e Y son iguales, pero su correlación es -1. Como ejemplo veanse los siguientes datos:

O_1	O_2
1	5
4	2
5	1
2	4
3	3

Si no se cumplen las condiciones de paralelidad la correlación no se consideraría una estimación del coeficiente de fiabilidad.

El incumplimiento de estas condiciones implica que no puede garantizarse que no haya habido efectos de factores que afectan las puntuaciones de la segunda medición.

El cuadro siguiente sintetiza algunas de las cuestiones en torno a la supuesta fuente de la variación de las puntuaciones y sus efectos sobre el coeficiente de fiabilidad.

Pone de manifiesto claramente como un instrumento puede ser perfectamente fiable y tener un coeficiente de correlación de Pearson distinto de 1, y, al revés, tener un coeficiente de correlación entre mediciones igual a 1 y no ser perfectamente fiable.

Efectos sobre la correlación de dos mediciones, efectuadas bajo el diseño de mediciones paralelas de la Teoría Clásica, de variaciones reales versus introducidas por el instrumento.

SUJETOS \ INSTRUMENTO DE MEDIDA	a. NO INTRODUCE NINGUNA VARIACION	b. INTRODUCE VARIACION LINEAL	c. INTRODUCE VARIACION ALEATORIA COMPATIBLE CON PARALELIDAD	d. INTRODUCE VARIACION NO LINEAL
1. NO SUFREN REALMENTE NINGUNA VARIACION	$r_{AB} = 1$	$r_{AB} = 1$	$r_{AB} < 1$	$r_{AB} < 1$
2. SUFREN VARIACION LINEAL	$r_{AB} = 1$	$r_{AB} = 1$	$r_{AB} < 1$	$r_{AB} < 1$
3. SUFREN VARIACION ALEATORIA COMPATIBLE CON PARALELIDAD	$r_{AB} < 1$	$r_{AB} < 1$	$r_{AB} < 1$	$r_{AB} < 1$
4. SUFREN VARIACION NO LINEAL	$r_{AB} < 1$	$r_{AB} < 1$	$r_{AB} < 1$	$r_{AB} < 1$

Estos coeficicientes podrían considerarse coeficientes de fiabilidad al poder satisfacerse los supuestos de paralelidad.

Instrumento perfectamente fiable.

11. Propuestas y dificultades en torno a las mediciones paralelas

El concepto de mediciones paralelas tiene tres funciones principales en TCT. Primero, fundamenta el concepto de puntuación verdadera que puede definirse mediante la esperanza a través de mediciones paralelas. Segundo, establece el puente necesario entre las ecuaciones de la TCT y su estimación. Tercero, resuelve la necesidad de obtener mediciones directamente comparables, un problema práctico importante por sí mismo.

Como Lord and Novick (1968, p. 48) señalan "para la mayoría de los propósitos la validez y la utilidad empírica del modelo descansa en el supuesto de independencia lineal y en la disponibilidad de mediciones paralelas". Además, el concepto de mediciones paralelas forma parte de muchos desarrollos teóricos en la TCT, por ejemplo el fundamento de la fórmula de Spearman-Brown, o la relación entre coeficiente alfa y coeficiente de fiabilidad. Por otra parte, el concepto de paralelidad está reapareciendo de diversos modos por razones teóricas y prácticas en las aproximaciones de la Teoría de la Respuesta al Item a la cuestión de la fiabilidad sin que se aborde la cuestión esencial de su contraste.

La imposibilidad de sostener la invariabilidad del objeto medido ha llevado al desarrollo de la teoría de las observaciones muestrales comparables como aproximación a la fiabilidad. Esta teoría sostiene que la estimación de la fiabilidad puede obtenerse como el grado de relación lineal entre dos conjuntos de puntuaciones suficientemente equivalentes. Siguiendo esta doctrina si dos medidas resultan suficientemente equivalentes entonces pueden utilizarse intercambiablemente y estimar la fiabilidad de las mismas mediante un estadístico de asociación.

Debido a que cada puntuación observada difiere de la verdadera por una variable aleatoria E, la TCT tiene que enfrentarse a la cuestión de la equivalencia entre mediciones observadas.

La TCT ha desarrollado una jerarquía de grados de equivalencia: 1) Mediciones con puntuaciones observadas idénticas; 2) Mediciones replicadas; 3) Mediciones paralelas; 4) Mediciones tau-equivalentes; 5) Mediciones esencialmente tau-equivalentes. Jöreskog (1971) resaltó un nivel 6), el de las mediciones congenéricas.

Las mediciones congenéricas presentan una relación lineal entre su puntuación verdadera y una variable latente lo que implica que las puntuaciones verdaderas de dos mediciones congenéricas mantienen una relación lineal entre ellas. Por ello podría considerarse que miden el mismo rasgo aunque posiblemente en diferente escala y con diferente error de medida.

Este conjunto de seis conceptos forma un escalograma de grados de equivalencia, con las mediciones congenéricas en el extremo más laxo pero con mayor probabilidad de realización empírica. Cualquier grado de equivalencia por debajo de las mediciones paralelas exige alguna clase de equiparación.

Como hemos visto, la correlación entre dos mediciones paralelas iguala el coeficiente de fiabilidad (Gulliksen, 1950/1983, pp. 13-14; Lord & Novick, 1968; p. 58). Esta demostración clásica es la piedra angular que permite la estimación de la fiabilidad en TCT. La demostración se basa en tres supuestos que se satisfacen si las medidas son paralelas, y dejan de satisfacerse si se desciende un peldaño más en el escalograma de equivalencia.

Debido a esta demostración, si dos mediciones son paralelas entonces su coeficiente de correlación (independientemente de su magnitud) es el coeficiente de fiabilidad de ambas. Todas las fórmulas de la teoría de la fiabilidad pueden estimarse si puede probarse que se dispone de mediciones paralelas, pero, desafortunadamente, las dos ecuaciones que definen mediciones paralelas no pueden ser directamente contrastadas y la disponibilidad de mediciones paralelas no está garantizada.

Estas dificultades han dado lugar a sucesivos intentos de esquivar las mediciones paralelas en la formulación teórica de la fiabilidad.

La teoría de las muestras estadísticamente equivalentes de Brown-Kelley es un intento de evitar estas dificultades formulando la definición de paralelismo en el plano observable. Gulliksen (1950) mostró que todas las fórmulas de la teoría de la fiabilidad pueden seguirse también de esta definición basada en observables, pero con dos limitaciones: 1) La correlación entre mediciones paralelas iguala el coeficiente de fiabilidad cuando el número de mediciones paralelas tiende a infinito, y 2) la misma equivalencia de la puntuación verdadera depende también de esta aproximación asintótica poco realista. En todo caso, esta formulación también requiere restringir la estimación de la fiabilidad a aquellas mediciones que satisfagan simultáneamente igualdad de medias, varianzas y correlaciones.

La teoría de la forma comparable de Tyron (1957) es otro intento de superar las dificultades anteriores esquivando la paralelidad. En esta formulación la estimación de la fiabilidad requiere que la segunda medición sea lo que denomina una forma comparable, es decir, que presente igual número de items, igual varianza media de los items, e igual covarianza media de los items. La estimación de la fiabilidad requiere satisfacer simultáneamente estas

condiciones cuya viabilidad práctica no es mucho mayor que la de las mediciones paralelas.

Lord y Novick (1968) son conscientes de que hay pocas probabilidades, si es que hay alguna, de que un psicólogo pueda encontrar tres mediciones que satisfagan estos criterios. Por ello introducen el concepto de mediciones nominalmente paralelas que no implica ninguna de las condiciones observables de paralelidad. La puntuación verdadera de un conjunto de mediciones nominalmente paralelas es la puntuación verdadera genérica, concepto implícito en la aproximación desarrollada por Cronbach, Rajaratnam y Gleser (1963). Sin embargo, no existe fundamento formal alguno para sostener que las mediciones nominalmente paralelas permitan estimar la fiabilidad (Lord and Novick, 1968, Ch. 8)

También se han producido algunos ensayos para resolver la cuestión del contraste de paralelidad.

Si tres o más medidas satisfacen las dos condiciones inobservables de paralelidad entonces necesariamente satisfacen las tres condiciones observables: medias iguales, varianzas iguales y correlaciones iguales. Basada en esta deducción, la doctrina Wilks-Votaw-Gulliksen de criterio estadístico de paralelidad provee un medio de contrastar la paralelidad a través de sus consecuencias observables mediante un estadístico chi-cuadrado (Gulliksen, 1950, Ch. 14). Desde el punto de vista de Gulliksen dos tests deben satisfacer el criterio estadístico y un criterio psicológico para poder considerarlos paralelos. Lord y Novick (1968) aceptaron esta doctrina para el contraste de la paralelidad, aunque su escasa viabilidad les indujo a introducir otros grados de equivalencia.

Jöreskog (1971) presentó un procedimiento general basado en ecuaciones estructurales para contrastar modelos de tests congenéricos, incluyendo como casos particulares las mediciones tau-equivalentes y las paralelas. Jöreskog resuelve la estimación de la fiabilidad a través de la relación con un factor latente con varianza fijada, sin supuestos de paralelidad adicionales.

No obstante subsisten un conjunto de dificultades que inducen a considerar que el contraste no esta plenamente resuelto.

Los criterios de formato y contenido psicológico no son suficientes para soportar el concepto de mediciones paralelas y por tanto la teoría clásica de la fiabilidad. Por eso es necesario disponer de un contraste de paralelidad. Si, efectuando el test de Wilks (1946), tres o más mediciones incumplen alguna de las condiciones observables de paralelidad, entonces es seguro que una o más de ellas no son mediciones paralelas y, por tanto, su correlación no puede considerarse el coeficiente de fiabilidad. Pero si tres o más mediciones satisfacen el test de Wilks, e incluso también el de Votaw, ello no garantiza

que sean mediciones paralelas por dos razones. Primero porque estos tests sólo establecen que las mediciones no difieren significativamente en los estadísticos observables, lo que no prueba su igualdad. Segundo, porque aunque se obtenga una igualdad exacta entre los estadísticos observables, en cuyo caso no es necesario el test de Wilks ni el de Votaw, ésta no garantiza el cumplimiento de las condiciones inobservables de paralelidad. Si una variable aleatoria toma sólo m valores discretos, entonces su distribución está completamente determinada por sus primeros m-1 momentos respecto al origen. Es obvio pues que la distribución de las puntuaciones verdaderas no puede determinarse por su media, varianza y correlaciones. Adicionalmente, la deducción que lleva de puntuaciones verdaderas iguales y varianzas de error iguales hasta las consecuencias observables no es reversible por lo que respecta a la condición de puntuaciones verdaderas iguales.

El número de puntos significativos en una escala afecta su precisión y, usualmente, es afectado por el número de items. Los tests paralelos requieren items con el mismo número de valores y tests con el mismo número de items.

El enfoque de los tests congenéricos permite estudiar la equivalencia lineal entre tests en diferentes escalas. Pero los tests congenéricos no retienen la idea de igual error de medida ni la idea de igual precisión, y por tanto, los tests congenéricos no pueden ser utilizados intercambiablemente.

Es obvio que si diferentes tests congenéricos presentan diferente error de medida también presentarán diferentes fiabilidades. El concepto de tests congenéricos no puede sostener la deducción del coeficiente de fiabilidad a partir del coeficiente de correlación. Dado que dos tests congenéricos difieren probablemente en medias y en varianzas, su correlación no puede verse como el coeficiente de fiabilidad.

El procedimiento de Jöreskog no fundamenta los métodos clásicos de estimación de la fiabilidad por lo que, aun si se dispone de un análisis factorial confirmatorio que justifique medidas congenéricas, la práctica de estimar la fiabilidad mediante estos procedimientos no queda amparada.

Por otra parte, el concepto de tests congenéricos se apoya en la estimación de un factor latente, pero la definición del factor latente se basa en la elección de los tests que resultan congenéricos. Diferentes conjuntos de tests considerados como congenéricos pueden producir diferentes estimaciones del factor latente, y, por tanto, diferentes estimaciones de la fiabilidad. Por ello es necesario definir la población de tests congenéricos para un constructo dado, y por ello, es necesario establecer un criterio de equivalencia más allá de que mantengan un conjunto de relaciones lineales no rechazadas por un modelo de ecuaciones estructurales.

La aproximación de Wilks-Votaw-Gulliksen considera las medias, las varianzas y las correlaciones, la de Jöreskog sólo no se ocupa de las medias. Ambas permiten rechazar la hipótesis de paralelismo bajo ciertas condiciones, pero ninguna de ellas permite establecer que dos o más mediciones son paralelas.

Las hipótesis referidas a la igualdad de estos estadísticos pueden rechazarse también mediante el análisis de estructuras de covarianza (modelos de ecuaciones estructurales), pero estos métodos tampoco permiten *probar* las condiciones inobservables de paralelidad.

Si los tests son paralelos el procedimiento de Jöreskog permite rigurosamente estimar la fiabilidad sin necesidad de utilizar los procedimientos tradicionales. Pero si los tests no son paralelos el método de Jöreskog continuara produciendo estimaciones de la fiabilidad en base a un hipotético rasgo latente común que precisamente viene definido por los tests cuyo grado de equivalencia está en discusión.

En mi opinión no se dispone de un test adecuado del concepto de mediciones paralelas y probablemente, dada la falta de reversibilidad de la cadena de deducciones entre inobservables y observables, ese test no pueda ser elaborado. Con ello el fundamento clásico de la estimación de la fiabilidad permanece especulativo aun disponiendo de varios tests para rechazar la paralelidad y de procedimientos alternativos para estimar la fiabilidad.

12. La confusión entre fiabilidad y precisión

Es frecuente en ciertos manuales de teoría clásica hablar de “precisión” como equivalente de “fiabilidad”. Quizás el origen de esta identificación entre fiabilidad y precisión esté en Lord y Novick (1968; pag. 134) cuando afirman: “La correlación entre mediciones verdaderamente paralelas tomadas de tal modo que la puntuación verdadera de la persona no cambie entre ellas es llamada frecuentemente *el coeficiente de precisión*” Y a continuación lo definen exactamente con la fórmula que define al coeficiente de fiabilidad (el cociente entre la varianza de las verdaderas y la varianza de las empíricas).

> Según la Espasa-Calpe, en la acepción que aquí nos interesa, se denominan “de precisión” aquellos “instrumentos construidos con singular esmero para obtener resultados exactos” pudiendo definirse la precisión como “la abstracción o separación mental que hace el entendimiento de dos cosas realmente identificadas, en virtud de la cual se concibe la una como distinta de la otra”. Según el “Diccionario de Términos Científicos y Técnicos” la precisión es “el grado en que se acerca el resultado de un cálculo o la lectura de un instrumento al valor de las medidas o cálculos que están libres de error” o bien, “la cualidad de ser exacto, o nítidamente o fijo”. Podría pensarse que este es un uso en castellano, pero en inglés -

el 'lenguaje fuente' para la psicometría y la ciencia en general- el campo semántico de "precisión" es el mismo. Así, el Collins, por ejemplo, define el término inglés "precision" como "the quality of being precise; accuracy. Characterized by or having a high degree of exactness" y "precise" como "strictly correct in amount or value. Using or opperating with total accuracy" Y, a su vez "accuracy" como "faithful measurement or representation of the truth." Resulta relevante como se define el término en inglés y en castellano en este asunto para reflejar que la cuestión que se discute aquí tiene que ver con la teoría psicométrica de fondo y no con un error de traducción.

A mi juicio la identificación de la precisión con el coeficiente de fiabilidad, que no es más que la correlación entre dos mediciones paralelas, resulta engañosa. Creo que puede establecerse una relación entre la fiabilidad de un instrumento (*definida como* grado en que el instrumento da siempre el mismo resultado bajo las mismas condiciones midiendo una misma magnitud que no ha variado) y la precisión (*definida como* grado en que el instrumento mide con exactitud y discierne entre los grados con fineza). Pero esa relación no puede sostenerse identificando fiabilidad y precisión especialmente cuando la fiabilidad se reduce a un coeficiente de correlación.

Creo que todo el mundo convendrá en un instrumento que mida la longitud con un margen de error de ± 1 mm. es menos preciso que otro que mida con un margen de error de $\pm 0'01$ mm. Aquí la precisión es la "agudeza" con la que el instrumento es capaz de "discernir" entre grados próximos de la magnitud. Esta concepción de la precisión es acorde a las definiciones de diccionario y, en psicometría, acorde con el significado del error típico de medida, una clase de variación o error promedio que se encuentra al medir con el instrumento.

Sin embargo esta concepción razonable de precisión es no solo distinta sino contradictoria con el concepto de fiabilidad definido como el grado de acuerdo entre dos mediciones con instrumentos de medida psicológicos evaluado con un coeficiente de correlación de Pearson. Esta definición de precisión es contradictoria con una concepción de la fiabilidad como estabilidad (test-retest) o consistencia (métodos de consistencia interna o formas o tests paralelos) entre mediciones.

En mi opinión, paradójicamente, los instrumentos de medición psicológicos (tests, escalas, cuestionarios) en la medida en que son precisos (discriminan con fineza grados de la variable) tienden a ser menos fiables (estables o consistentes), y en la medida en tienden a ser más fiables probablemente están siendo menos precisos.

Por ejemplo, en pruebas de actitudes o de personalidad cuanto más globales, generales y semejantes entre si son las preguntas más probable es que las respuestas de los sujetos a las mismas se mantengan estables o consistentes. En la medida en que las preguntas se refieran a aspectos de detalle, especifiquen o se acerquen a

conductas concretas, o a matices de actitud, es más probable que las respuestas a las mismas de los sujetos estén expuestas a variaciones -genuinas o indeseadas desde la medición- que pueden repercutir en una menor correlación entre mediciones.

Aunque no es un ejemplo de fiabilidad de instrumentos de medida, la contradicción entre fiabilidad (definida como correlación) y precisión aparece muy claramente en la cuestión de la fiabilidad interjueces. Supongamos que dos jueces distintos observan a un grupo de sujetos en un contexto determinado durante un número N de sesiones. Los observadores llevan una hoja de registro y cuentan cuántas veces se producen determinados eventos. Cuanto más fina sea la discriminación (precisión) que tienen que hacer los jueces menos fiabilidad interjueces (medida por la correlación de Pearson entre sus conteos para N observaciones) encontraremos. Cuanto menos fina sea la discriminación a hacer mayor facilidad para el acuerdo. Por ejemplo, es más fácil que dos "ojeadores" de fútbol nos clasifiquen con acuerdo (mayor fiabilidad como correlación) un conjunto de candidatos en "buenos" y "no tan buenos", que que nos clasifiquen con acuerdo ese conjunto de candidatos en su orden de calidad señalando el primero, el segundo...(menor fiabilidad como correlación). (Una cuestión paralela pero distinta es que cuanto más "objetivo", menos expuesto a discusión, más claro y menos ambiguo aquello que hay que observar mayor será el acuerdo).

Un instrumento incapaz o insensible para detectar cambios en los sujetos en determinado periodo de tiempo aparecerá como un instrumento fiable con una alta estabilidad temporal. Un instrumento incapaz de obtener matices y aspectos diversos de pregunta a pregunta presentará una alta consistencia interna. Este tipo de instrumento se asemeja a un gran paquidermo que es incapaz de notar las irregularidades sutiles del suelo simplemente porque las aplasta el mismo.

Si hiciéramos un instrumento de medida para detectar donde residen los sujetos formado por las cuestiones:

¿En que planeta reside Ud.? y

¿En que continente reside?

los resultados serían más estables y consistentes de medición en medición que otro formado por las preguntas:

¿En que Estado reside Ud.?

¿En que Nación reside Ud.?

y este aún sería más fiable para la teoría clásica que otro formado por las preguntas:

¿En que ciudad reside Ud.?

¿En que calle vive Ud.?

Cuanto más específicas las preguntas mayor probabilidad de detectar variación y, consecuentemente, mayor probabilidad de que la correlación entre mediciones sucesivas, entre partes de la misma medición o entre items sea más baja. La precisión y la supuesta "fiabilidad" de la teoría clásica no sólo no son la misma cosa sino que son cosas contradictorias.

El lector argumentará que en el ejemplo anterior nosotros podríamos comprobar objetivamente donde vive el sujeto y por tanto distinguir el cambio real (la verdadera ubicación del sujeto) y la variación debida al instrumento, que, en todos los casos, podría ser igual de buena. Efectivamente. Pero el problema reside justamente en que en psicología *no* tenemos un modo objetivo (distinto de instrumentos de medición psicológica o de especulaciones subjetivas más o menos fundadas) para ver cual es de verdad el grado en que los sujetos poseen cualidades como inteligencia, atención, actitud xenófoba, satisfacción, psicoticismo o neuroticismo. El problema reside en que no accedemos a esos inobservables sino es a través del instrumento y que, la teoría clásica, ni nos permite acceder independientemente al inobservable ni nos permite, por tanto, estimar la precisión.

Quizás el lector pueda pensar que las preguntas del ejemplo anterior acerca de planetas y continentes son exageradas, demasiado generales, demasiado burdas. A mi me parece que representan justamente muchas preguntas al uso de tests de actitudes, de personalidad o de inteligencia que son demasiado generales, demasiado imprecisas, demasiado ambiguas, demasiado burdas, y quizás, por todo ello razonablemente "fiables".

Si el lector revisa cuidadosamente el contenido de tests bien conocidos, podrá encontrar muchas *preguntas sobre "en qué planeta vive el sujeto"*. De muestra véanse estos tres items del 16PF:

> *"¿Cuáles (sic) de las siguientes palabras es diferente a las otras dos? A. Algo; B.Nada; C. Mucho."*

> *"Poseo suficiente energía para enfrentarme a todos mis problemas A. Siempre; B. Frecuentemente; C.Raras veces."*

> *"Cuando me critican duramente por algo que no he hecho: A. No me siento culpable; B. Término medio; C. Todavía me siento un poco culpable."*

La primera pregunta se pretende que mida inteligencia -y la respuesta correcta de una inteligencia sana es la B, al

parecer-; las otras dos miden diversos factores de personalidad más o menos bien definidos (en concreto, "poca-mucha fuerza del ego" y "adecuación imperturbable-tendencia a culpabilidad", respectivamente). El resto del test es parecido.

He citado intencionalmente un test bien conocido y reconocido, porque el punto de vista que estoy expresando afecta de diversas formas a la mayoría de los tests. Otros cuestionarios y tests bien conocidos presentan preguntas cuyos propósitos, formato y enfoque no difieren sustancialmente de estos a efectos de esta discusión.

A mi juicio la *precisión* está relacionada con el grado en que un test escala adecuadamente la variable a medir. Escalar adecuadamente implica que el test es capaz de (1) medir la variable en todo su rango observable (desde su mínimo real hasta su máximo real observable), (2) representando adecuadamente todas las zonas intermedias, y, (3) discerniendo grados de presencia de la variable tan pequeños como sea posible. Es decir, un test es preciso en la medida en que sea capaz de discernir entre estados de la variable cuya diferencia d sea menor que ε, para un ε tan pequeño como se quiera. Para un test dado el rango de la escala es un estimador burdo de la precisión. Probablemente, este estimador es inversamente proporcional a ε.

No obstante una cuestión esencial es establecer la naturaleza teórica de la variable a medir. Es elegante matemáticamente suponer un rasgo latente continuo, y cómodo operativamente utilizar un test que "mide" de un modo discreto. Ni una ni otra atribución se fundan en una teoría psicológica fundamentada sobre la naturaleza de la variable a medir. La cuestión sobre cuan precisa puede ser la medida en psicología exigiría una teoría psicológica que explicara la naturaleza de la variable y cuantos grados de discernimiento pueden establecerse en ella.

Un tratamiento de la precisión desde este enfoque implica la discusión del grado en que el discernimiento entre valores adyacentes en la escala (1) refleja diferencias reales en grado -y no meramente inducidas por el instrumento, aunque sea de forma constante-, y (2) se mantiene constante a través de diversas zonas de la escala, y (3) se mantiene constante a través de mediciones (concepción clásica de fiabilidad como constancia) y (4) de otros universos de generalización, como sujetos, condiciones, etc.

Estas reflexiones llevan más allá del problema de la paradoja entre fiabilidad y precisión, insinúan una linea de estadísticos de precisión (y también de fiabilidad) diferente de la clásica y cuestionan otras características de los tests,

aspectos prácticos y teóricos en cuyo comentario, necesariamente extenso, no entraremos ahora.

Si se presentan juntos todos los razonamientos críticos que he ido desarrollando a lo largo del capítulo acerca de la teoría clásica de tests y, muy en particular, acerca de la teoría de la fiabilidad, parece poco probable que ésta pueda sostenerse.

En buena parte estas críticas pueden extenderse a los modelos basados en consistencia interna que veremos en capítulos siguientes. Han sido presentadas ahora porque para su desarrollo básico es suficiente tomar en cuenta las cuestiones centrales de interpretación, diseño y modelo subyacente que ya han sido expuestas hasta este punto.