Psicothema
2000. Vol. 12, Supl. nº 2, pp. 386-388

ISSN 0214 - 9915 CODEN PSOTEG

La cuestión de las mediciones paralelas

Josep L. Melià
Universidad de Valencia

Las medidas que tienen puntuaciones verdaderas identicas y errores linealmente independientes con varianzas iguales se denominan medidas paralelas (Lord & Novick, 1968). Desde el origen de la Teoría Clásica de Tests el concepto de mediciones paralelas ha jugado un papel central como piedra angular del concepto y la estimación de la fiabilidad (Gulliksen, 1950; Lord & Novick, 1968). Las dificultades para encontrar mediciones comparables tan semejantes fue pronto conocida proponiéndose otros conceptos que expresan diferentes grados de comparabilidad. El propósito de estos conceptos de comparabilidad relajada fue encontrar el mayor nivel de viabilidad empírica todavía capaz de sustentar la base teórica del coeficiente de fiabilidad. Aunque Wilks (1946) y Jöreskog (1971) aportaron métodos débiles para determinar si dos medidas son paralelas, el concepto clásico de mediciones paralelas permanece no testado y su posibilidad empírica no está clara. El propósito de este trabajo es trazar el desarrollo del concepto de mediciones paralelas como fundamento de la fiabilidad y discutir algunas de sus limitaciones.

The concept of parallel measurements. Measures that have identical true scores and linearly experimentally independent errors that have equal variances are named parallel measurements (Lord & Novick, 1968). Since the origin of the Classical Test Theory (CTT) the concept of parallel measurement has played a central role as a cornerstone of the concept and the estimation of reliability (Gulliksen, 1950; Lord & Novick, 1968). The difficulties in finding such similar comparable measurements were early known, and some other concepts, wording different levels of comparability, were proposed. The purpose of these relaxed concepts of comparability was to find a higher level of empirical availability yet capable of sustaining a theoretical base for the reliability coefficient. Although Wilks (1946) and Jöreskog (1971) provided soft methods to determine if two measures are parallel tests, the classical concept of parallel measurements remains untested and its empirical possibility is far from clear. The aim of this paper is to trace the development of the concept of parallel measurements as a base for reliability and discuss some of its limitations.

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.

Correspondencia: Josep L. Meliá
Facultad de Psicología
Universidad de Valencia
46010 Valencia (Spain)
E-mail: Jose.L.Melia@uv.es

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. 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 $E_{ga} = X_{ga} - \tau_{ga}$, 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: $\tau_{ga} \equiv \mathcal{E}(X_{ga}) = \mathcal{E}(X_{g'a})$ y $F(E_{ga}) = F(E_{g'a})$. 3) Mediciones paralelas: $\tau_{ga} = \tau_{g'a}$; y $\sigma^2(E_{ga}) = \sigma^2(E_{g'a})$. 4) Mediciones tau-equivalentes: $\tau_{ga} = \tau_{g'a}$. 5) Mediciones esencialmente tau-equivalentes: $\tau_{ga} = \tau_{g'a} + a_{gg'}$. 6) Las mediciones congenéricas presentan una relación lineal entre su puntuación verdadera y una variable latente (Jöreskog, 1971), lo que implica $\tau_{ga} = b_{gg'}\tau_{g'a} + a_{gg'}$, de modo que 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.

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 (Meliá, 1993). 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.

Evitando las Mediciones Paralelas

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. 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 no es mayor.

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)

Contrastando la 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.

Dificultades de un Contraste No Resuelto

La disponibilidad empírica de mediciones paralelas y por tanto, el contraste de la paralelidad, constituyen un pilar central de la TCT. Debe tenerse en cuenta que el coeficiente de fiabilidad no prueba o expresa ningún grado de equivalencia. Si dos medidas son paralelas su correlación es el coeficiente de fiabilidad, sea cual sea su magnitud. Pero si dos medidas no son paralelas su correlación no es el coeficiente de fiabilidad incluso si su correlación es perfecta (Melià, 1993).

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. No se dispone de un test adecuado del concepto de mediciones paralelas y probablemente, dada la no-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 y la aplicación de la misma teóricamente no justificada.

Referencias

Cronbach, L.J., Rajaratnam, N. y Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology, 16,* 137-163.

Gulliksen, H. (1950/1987). *Theory of mental tests.* Hillsdale: Lawrence Erlbaum Associates.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric test. *Psychometrika, 36,* 109-133.

Lord, F.M. y Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Melià, J.L. (1993) *Apuntes sobre Teoría Clásica de Tests.* Valencia. Cristobal Serrano.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54,* 229-249.

Wilks, S.S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics 17,* 257-281.