Psicothema 2009. Vol. 21, nº 2, pp. 321-325
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Implementación de procedimientos gráficos y analíticos para la construcción de formas paralelas

Pere Joan Ferrando Piera, Urbano Lorenzo-Seva y Rafael Pallero González*
Universidad Rovira i Virgili: Centro de Investigación y Medida de la Conducta. Departamento de Psicología y * ONCE

Se presenta un trabajo de instrumentación en el que se implementan procedimientos gráficos y objetivos para construir formas paralelas a partir de un conjunto de ítems calibrados mediante la teoría clásica de los tests. El procedimiento gráfico es el de subtests apareados propuesto por Gulliksen. El procedimiento objetivo se basa en los criterios propuestos por van der Linden y Boekkooi-Timminga y utiliza programación cero-uno. El programa FOR-PAR, de fácil uso y libre distribución, es auto-contenido y permite obtener las formas paralelas a partir de las puntuaciones de los ítems. FOR-PAR cubre una necesidad aplicada importante ya que la necesidad de elaborar formas paralelas surge con relativa frecuencia en la práctica y, hasta ahora, no existía ningún programa de este tipo. La aplicación de los procedimientos implementados y el uso del programa se ilustran mediante un estudio empírico basado en un test de ajuste a la discapacidad visual.

Implementing graphical and analytical procedures for developing parallel tests. We present an instrumental study in which two general procedures (graphical and objective) for constructing parallel forms are implemented. Both procedures are based on an item pool which is calibrated by means of the classical test theory. The graphic procedure is Gulliksen's matched random subtests method. The objective procedure is based on the criteria proposed by van der Linden and Boekkooi-Timminga, and uses zero-one programming. The stand-alone program FOR-PAR is free and user-friendly, and allows the parallel forms to be obtained directly from the item scores. FOR-PAR covers an important need in applied research, because developing parallel forms is a requirement in some applications, and, so far, no programmes of this type were available. The procedures which are implemented were applied in an illustrative example based on an adjustment test intended for visually handicapped people.

A pesar del dominio de los modelos de teoría de respuesta al ítem en las últimas décadas, la teoría clásica del test (TCT) sigue utilizándose ampliamente en aplicaciones donde se requieren soluciones sencillas a problemas prácticos de medida o donde debe trabajarse con muestras reducidas (e.g., Muñiz, 1992). Desde este punto de vista práctico, cabe decir que la mayor parte de los procedimientos derivados de la TCT que se requieren en investigación aplicada (particularmente el análisis de ítems o la estimación de la fiabilidad) se encuentran implementados en programas comerciales de amplia difusión. Sin embargo, curiosamente, no lo están algunos procedimientos de indudable interés e incluso necesidad.

La construcción de tests que satisfagan lo mejor posible las condiciones de estricto paralelismo a partir de un conjunto calibrado de ítems es un proceso laborioso y de claro interés aplicado. La correlación entre dos formas paralelas es el procedimiento genuino para evaluar la fiabilidad del test que se deriva de los principios de la TCT (Muñiz, 1992). En la práctica, sin embargo, las formas paralelas no se usan tanto para evaluar la fiabilidad como para otros propósitos. Así, en situaciones clínicas, educativas o laborales suele tener que evaluarse a los participantes más de una vez, y es conveniente hacerlo con instrumentos que proporcionen medidas equivalentes pero que estén formados por ítems distintos (e.g., Garaigordobil, 2004). De esta forma se pueden reducir potenciales efectos de falseamiento o práctica (e.g., Martínez-Cardeñoso, Muñiz y García-Cueto, 2000). El proceso se requiere también cuando se quieren desarrollar versiones reducidas de un test, de manera que las formas resultantes tengan la máxima equivalencia posible (e.g., Pino et al., 2006).

Las directrices generales para la obtención de formas paralelas han sido discutidas en detalle por Thorndike (1950). De acuerdo con ellas, Gulliksen (1950) propuso un procedimiento gráfico relativamente simple que puede ser considerado como el procedimiento estándar. Dicho procedimiento, sin embargo, tiene un fuerte componente de subjetividad, no permite formular un criterio inequívoco de asignación de los ítems y, en muchas aplicaciones, da lugar a resultados subóptimos. Para superar estas limitaciones van der Linden y Boekkooi-Timinga (1988) propusieron un procedimiento basado en un criterio objetivo e implementado en un algoritmo derivado de las técnicas de programación cero-uno. El procedimiento representa una clara mejora con respecto al método básico de Gulliksen y, desde el punto de vista metodológico, fue

Fecha recepción: 12-6-08 • Fecha aceptación: 7-10-08
Correspondencia: Pere Joan Ferrando Piera
Facultad de Psicología - URV
Carretera de Valls, s/n
43007 Tarragona (Spain)
E-mail: perejoan.ferrando@urv.cat

totalmente resuelto por sus autores. Sin embargo, van der Linden y Boekkooi-Timinga no implementaron ningún programa específico y se limitaron a mostrar que el procedimiento podía aplicarse utilizando un programa general de minimización. Es dudoso que un investigador aplicado en Psicología se decida a adaptar algoritmos de programación para desarrollar formas paralelas de un test, y no es ninguna sorpresa que los autores hayamos sido incapaces de encontrar una sola investigación aplicada en la que se haya utilizado el procedimiento objetivo.

El presente artículo presenta y describe un programa de fácil uso y libre distribución para obtener formas paralelas a partir de un conjunto calibrado de ítems. El programa permite utilizar el procedimiento de Gulliksen a partir de la representación gráfica de los ítems y obtiene la solución objetiva de acuerdo con los criterios de van der Linden y Boekkooi-Timinga descritos más abajo. En la siguiente sección se describen los fundamentos metodológicos de los procedimientos implementados. A continuación se describe el programa. Finalmente, se ilustran los procedimientos mediante un ejemplo empírico.

Descripción y justificación de los procedimientos

Considérese un test formado por n ítems. Sea $X= X_1 +\ldots+X_n$ la puntuación total obtenida como suma simple de las puntuaciones en los ítems. Como índice de dificultad de un ítem j se tomará su puntuación media, y como índice de discriminación la correlación entre las puntuaciones de dicho ítem y las puntuaciones totales en el test: r_{jX} . Los dos resultados básicos para los procedimientos implementados son (e.g., Gulliksen, 1950):

$$\overline{X} = \sum_{j}^{n} \overline{X}_j \tag{1}$$

y

$$s_x = \sum_{j}^{n} r_{jX} s_j \tag{2}$$

donde s_X es la desviación típica del test total. En el caso particular de ítems binarios, los resultados se reducen a:

$$\overline{X} = \sum_{j}^{n} p_j \tag{3}$$

$$s_X = \sum_{j}^{n} r_{jX} \sqrt{p_j\left(1-p_j\right)} \tag{4}$$

donde p_j es la proporción de respuestas puntuadas como 1 (índice de dificultad). Las ecuaciones (3) y (4) indican que, en el caso binario, los índices de dificultad y discriminación de los ítems determinan totalmente la media y varianza que tendrán las puntuaciones en el test total. En el caso de ítems más continuos, la determinación de la varianza es más aproximada ya que también influyen las varianzas de los ítems (véase 2).

Supóngase que se dispone de un conjunto de ítems calibrado en una muestra grande y representativa, de forma que las estimaciones de dificultad y discriminación son estables. Se pueden entonces construir formas paralelas: (a) formando parejas de ítems que sean similares en dificultad y discriminación, y (b) asignando un miembro de cada pareja a cada una de las dos formas a desarrollar. Con este procedimiento, y de acuerdo con las relaciones descritas arriba, las dos formas tendrán medias y varianzas aproximadamente iguales. Éstas son las condiciones necesarias de paralelismo utilizadas en la práctica (Lord y Novick, 1968).

En el procedimiento de Gulliksen (1950) cada ítem se representa en un gráfico bivariado como un punto cuyas coordenadas son el índice de dificultad y el índice de discriminación. Las parejas de ítems se forman por inspección, uniendo ítems que se encuentran cerca entre ellos en el plano. Finalmente, una vez decididas las parejas, los miembros se asignan al azar a cada forma. Un típico gráfico de Gulliksen, el correspondiente al ejemplo ilustrativo de este artículo, puede verse en la figura 1. Algunas parejas de ítems parecen obvias, por ejemplo, 12 y 20. Otras no lo son tanto. Así, por ejemplo, no está claro cómo deben emparejarse los ítems del cluster: 6, 8, 16, 3 y 11, y cada elección condicionará por supuesto los siguientes emparejamientos. Esta falta de un criterio objetivo de asignación es la principal debilidad del método. Debilidad que, obviamente, se hará más patente cuanto mayor sea el conjunto de ítems y más compleja la estructura de los puntos.

Para obtener un criterio objetivo de asignación, van der Linden y Boekkooi-Timinga (1988) consideraron como medida básica la distancia euclídea entre dos puntos. Dados los ítems j y k, la distancia entre ellos es:

$$\delta_{jk} = \left[\left(\overline{X}_j - \overline{X}_k\right)^2 + \left(r_{jX} - r_{kX}\right)^2\right]^{/2} \tag{5}$$

Sea ahora I_{jk} una variable indicadora que toma los valores: 1 si los ítems j y k forman pareja, y 0 en otro caso. El criterio se puede definir operativamente como sigue: minimizar

$$\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} \delta_{jk} I_{jk} \tag{6}$$

Sujeto a la restricción:

$$\sum_{j=1}^{k-1} I_{jk} + \sum_{j=k+1}^{n} I_{kj} = 1 \tag{7}$$

Conceptualmente el criterio (6) significa minimizar la suma de las distancias entre las parejas de ítems resultantes. La restricción (7) garantiza que para cada ítem la variable indicadora toma el valor 1 una sola vez. Esto implica que cada uno de los ítems sólo puede aparecer en una de las parejas.

La minimización de (6) sujeto a la restricción (7) es un problema típico de programación lineal cero-uno. Existen bastantes programas de tipo general que implementan este tipo de programación, y que pueden adaptarse a este problema concreto. Van der Linden y Boekkooi-Timinga (1988) utilizaron el programa LANDO basado en un algoritmo propuesto por Land y Doig (1960). En el presente trabajo se ha desarrollado una versión simplificada del algoritmo propuesto por Balas (1965). Esta simplificación se ha conseguido implementando las restricciones específicas del problema en lugar de las restricciones generales consideradas en el algoritmo de Balas. Se verificó que nuestro algoritmo simplificado y el algoritmo general de Balas llegaban al mismo resultado. Sin embargo, la mayor simplicidad de la adaptación hace que el programa sea considerablemente más rápido y eficiente, lo que es de interés sobre todo en conjuntos grandes de ítems.

El programa FOR-PAR: descripción y disponibilidad

El programa FOR-PAR se ha desarrollado como una aplicación EXCEL de Microsoft que está firmada digitalmente para facilitar su utilización en el entorno Windows. Se trata de un programa auto-contenido en el sentido de que no requiere de ningún tipo de resultados previos obtenidos con otros programas. Como input la aplicación requiere: (a) la matriz de puntuaciones, de dimensión *N* participantes × *n* ítems, (b) el número de participantes, (c) el número de ítems, y (d) el nivel de significación deseado para los intervalos de confianza de medias y desviaciones. Si el número de ítems es impar, la aplicación selecciona el mayor número posible de parejas y descarta el ítem que queda más distanciado del resto.

A partir de los datos proporcionados, FOR-PAR calcula las puntuaciones totales en el test, los índices de dificultad y de discriminación y la matriz de distancias inter-ítem (de dimensión $n \times n$). Dicha matriz es el input para el algoritmo de minimización modificado de Balas, el cual extrae sucesivamente las *n/2* parejas de ítems y asigna aleatoriamente cada uno de los miembros a la forma A o a la forma B.

El output del programa consiste en el gráfico de Gulliksen, la composición de las formas A y B, y los descriptivos correspondientes a las puntuaciones en las dos formas: medias, desviaciones típicas y correlación entre ellas. FOR-PAR calcula también los intervalos de confianza correspondientes a las medias y desviaciones al nivel de confianza especificado por el usuario. Con respecto al gráfico, la aplicación marca con diferente color los ítems asignados a A y B, lo que facilita la inspección visual del resultado obtenido de acuerdo con el criterio objetivo. El usuario puede también utilizar este gráfico para decidir de forma subjetiva la formación de parejas y la asignación de los ítems a las formas.

FOR-PAR es de libre distribución y está disponible gratuitamente para todos los lectores que lo deseen. Para ello, deben solicitarlo por e-mail o correo a uno de los dos primeros firmantes del artículo y se les facilitará el programa y un breve manual de utilización.

Ejemplo ilustrativo

El Cuestionario Tarragona de Ansiedad para Ciegos (CTAC; Pallero, Ferrando y Lorenzo, 2006) es un instrumento bidimensional, utilizado sobre todo en adultos que han perdido total o parcialmente la visión. El test tiene 35 reactivos en formato Likert de 5 puntos, y sus dos subescalas: Ansiedad Cognoscitiva (AC) y Ansiedad Fisiológica muestran propiedades psicométricas más que aceptables. Sin embargo, la experiencia indica que en algunos pacientes la administración del cuestionario completo resulta fatigosa. Además, uno de los usos del CTAC es el de evaluar la eficacia de los programas de reducción de ansiedad, y esto requiere tomar medidas antes y después. Por estas razones, se consideró de interés elaborar dos versiones equivalentes del test, cada una de ellas conteniendo la mitad de los ítems del test total. En este ejemplo consideraremos solamente la escala AC. Está formada por 20 reactivos, y en estudios previos la fiabilidad de sus puntuaciones se estimó en 0.88 (alfa de Cronbach). El objetivo es el de desarrollar dos mitades paralelas de la escala. De acuerdo con la predicción de Spearman-Brown, la correlación entre las mitades resultantes debería ser, aproximadamente, de 0.78, una fiabilidad que aún se consideraría aceptable.

La figura 1 muestra el gráfico bivariado de Gulliksen tal como lo devuelve el output de FOR-PAR. Los cálculos se desarrollaron a partir de las puntuaciones de una muestra de 352 participantes en los 20 ítems. Si el usuario quisiera utilizar el método de Gulliksen, entonces debería elegir las 10 parejas de ítems mediante inspección visual.

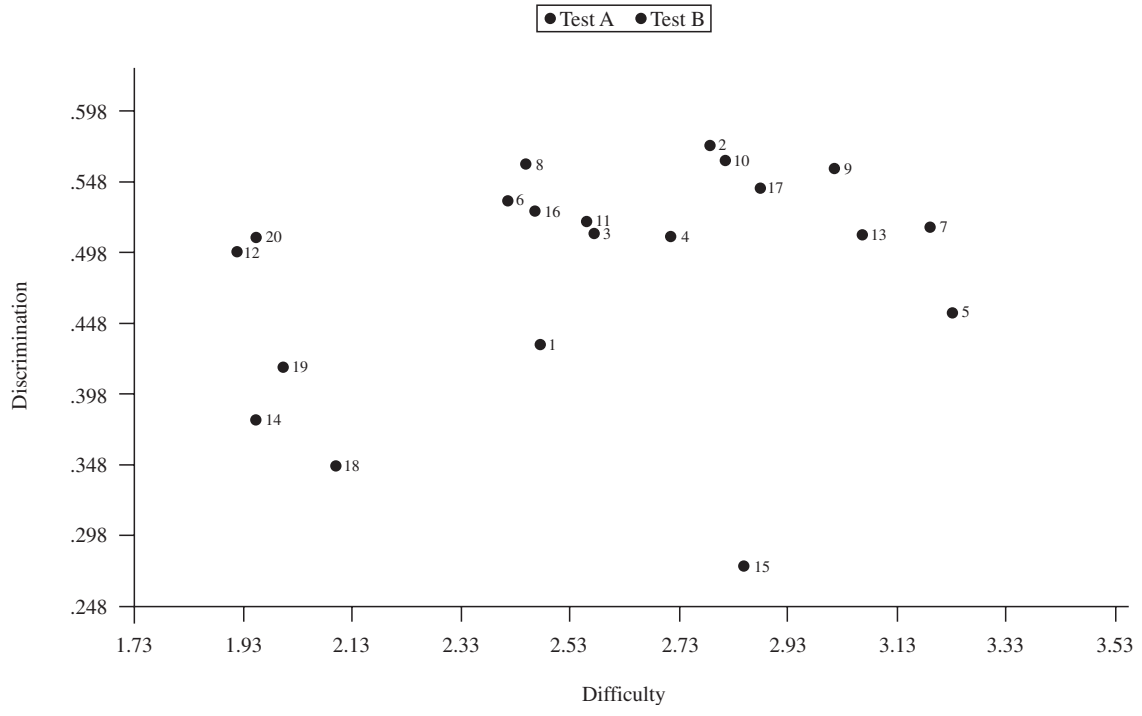

Figura 1. *Gráfico de Gulliksen*

El resultado de la aplicación del criterio objetivo, con las parejas de ítems seleccionadas en cada paso y los ítems asignados a la forma A y a la forma B, se presenta en la tabla 1. Si se relacionan las parejas formadas objetivamente con el gráfico de la figura 1 se observa que la asignación es bastante razonable. Debe tenerse en cuenta, sin embargo, que las parejas que se van determinando son cada vez menos similares, por lo que, excepto en configuraciones muy favorables, las últimas parejas pueden ser relativamente distintas. Éste sería el caso de la última pareja en la tabla 1. De hecho, la figura 1 sugiere que el ítem 15 es prácticamente un outlier.

Hay un segundo resultado que conviene también discutir con respecto a la solución alcanzada en la tabla 1. En el caso binario, tanto las dificultades como las discriminaciones se mueven en un rango efectivo de 0 a 1. Por tanto, de acuerdo con el criterio de la ecuación (5) ambos índices tienen el mismo peso en la determinación de las distancias. En el caso Likert, sin embargo, el rango de las dificultades (medias) es considerablemente mayor que el de las discriminaciones. Las dificultades tienen pues un peso mayor, y el criterio lleva sobre todo a minimizar diferencias entre medias. Si se desea que ambos índices tengan, como en el caso binario, el mismo peso, la solución más simple sería reescalar las dificultades en el intervalo 0-1.

La tabla 2 muestra las medias y desviaciones típicas de las dos formas junto con sus correspondientes intervalos de confianza al 95%. En ambos casos las correspondientes estimaciones puntuales son bastante similares (sobre todo las de las desviaciones), y los intervalos de confianza se solapan. La correlación entre las mitades se estimó en r_{ab}= 0.79. Si aceptamos que se cumplen los supuestos de paralelismo, esta correlación se interpreta como la fiabilidad de las puntuaciones de cada una de las dos mitades. De ser así, el resultado se ajusta a la fiabilidad que se había predicho inicialmente a partir de la fórmula de Spearman-Brown (r_{ab}= 0.79). En conjunto, parece que el objetivo de construir dos mitades paralelas de la escala AC se ha conseguido.

Tabla 1
Determinación objetiva de las parejas de ítems y asignación a las formas A y B

Pareja	Forma A	Forma B
1ª	3	11
2ª	10	2
3ª	12	20
4ª	8	16
5ª	19	14
6ª	13	9
7ª	7	5
8ª	1	6
9ª	17	4
10ª	18	15

Tabla 2
Medias y desviaciones típicas de las mitades A y B

	Forma A	Forma B
Media	25.24	26.18
I. de C. al 95%	(24.37; 27.01)	(25.32; 27.20)
Desviación típica	8.38	8.27
I. de C. al 95%	(7.81; 9.07)	(7.71; 8.98)

Discusión

En este artículo se presenta un trabajo de instrumentación en el que se implementan procedimientos para construir formas paralelas. Con respecto a contribuciones anteriores, nuestro estudio se basa en algoritmos de programación más modernos y simples para alcanzar los criterios de minimización y llegar a una solución objetiva. Sin embargo, consideramos que la principal contribución del trabajo no es la innovación, sino la de cubrir una necesidad en psicometría aplicada, ya que no existen programas similares al que aquí se presenta.

Aunque no sea una limitación del método ni del programa, debería quedar claro que el principal problema aplicado que puede surgir cuando se utilizan los procedimientos aquí implementados sería el que los anglosajones denominan como «capitalization on chance». Para que el método proporcione resultados válidos y generalizables, las estimaciones de los parámetros de los ítems deben estar basadas en muestras grandes y representativas. Si las estimaciones no son estables, es decir, si los índices pueden fluctuar mucho a través de diferentes muestras, entonces la solución obtenida cumplirá aceptablemente las condiciones de paralelismo en la muestra en la que ha sido obtenida. Sin embargo, posiblemente no lo hará en estudios de validación cruzada cuando se trabaje con nuevas muestras. De hecho, en caso de estimaciones muy inestables obtenidas en muestras muy pequeñas, los métodos tradicionales basados en la asignación de ítems pares e impares podrían llegar a resultados más generalizables. Con respecto a este punto, el grupo de calibración considerado en el ejemplo empírico sería quizás aún pequeño para los fines del procedimiento.

Por último, debe tenerse en cuenta que el procedimiento objetivo se basa en criterios puramente estadísticos y no tiene en cuenta el contenido de los ítems. Por tanto, no puede garantizarse que la solución objetiva lleve a un contenido equilibrado en ambas formas (i.e., validez de contenido). Como se ha apuntado anteriormente, el investigador puede utilizar la información gráfica y asignar subjetivamente los ítems a las formas, lo cual le permite tener en cuenta los criterios de contenido. Sin embargo, parece mejor opción trabajar sobre la solución objetiva introduciendo, si es posible, las modificaciones pertinentes. Así, por ejemplo, se puede tomar la lista de parejas proporcionada por el programa y decidir la asignación de cada miembro a una u otra forma de manera que se equilibren los contenidos de los ítems.

Agradecimientos

Esta investigación se ha realizado con el apoyo de la ONCE (contrato T07067S).

Referencias

Balas, E. (1965). An additive algorithm for solving linear programs with zero-one variables. *Operations Research*, *13*, 517-546.

Garaigordobil, M. (2004). Intervención psicológica en la conducta agresiva y antisocial con niños. *Psicothema, 16*, 429-435.

Gulliksen, H. (1950). *Theory of mental tests*. New Tork: Wiley.

Land, A.H., y Doig, A. (1960). An automatic method of solving discrete programming problems. *Econometrika*, *28*, 497-520.

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

Martínez-Cardeñoso, J., Muñiz, J., y García-Cueto, E. (2000). Mejora de las puntuaciones de los tests mediante entrenamiento. *Psicothema, 12*, 363-367.

Muñiz, J. (1992). *Teoría clásica de los tests*. Madrid: Pirámide.

Pallero, R., Ferrando, P.J., y Lorenzo, U. (2006). *Cuestionario Tarragona de Ansiedad para Ciegos.* Madrid: ONCE.

Pino, O., Guilera, G., Gómez, J., Rojo, J.E., Vallejo, J., y Purdon, S.E. (2006). Escala breve para evaluar el deterioro cognitivo en pacientes psiquiátricos. *Psicothema, 18*, 447-452.

Thorndike, R.L. (1950). Reliability. En E.F. Lindquist (ed.): *Educational measurement*. Washington: American Council on Education.

van der Linden, W.J., y Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subtests method. *Applied Psychological Measurement, 12*, 201-209.