Psicothema
2000. Vol. 12, nº 2, pp. 303-307

ISSN 0214 - 9915 CODEN PSOTEG

Estimación de habilidad mediante ítems isomorfos. Efectos en la fiabilidad de las puntuaciones

Javier Revuelta
Universidad Autónoma de Madrid

En este artículo se aborda el problema de la imprecisión en los parámetros de los ítems y su efecto las puntuaciones de los sujetos en los tests adaptativos. En particular se considera la imprecisión introducida en el test por el proceso de incluir ítems isomorfos en el banco. La investigación se lleva a cabo mediante un estudio de Monte Carlo en el que la precisión se calcula bajo diferentes niveles de error en los parámetros de los ítems. Los resultados indican que el proceso de creación de isomorfos puede ser una alternativa viable, pero es necesario previamente obtener una estimación del error introducido por dicho proceso.

Estimating ability from items isomorphs. effects on the reliability of the test scores. This article focuses on the errors of the item parameter estimates and their effect on the reliability of the test scores. In particular we consider the errors introduced by the process of creating item isomorphs. The research is conducted by means of a Monte Carlo simulation. The simulation includes several conditions regarding the size of the errors and the number of isomorphs in the item pool of an adaptive test. The results show that the errors due to the isomorphing process do not compromise the psychometric status of the test scores except in the condition with highest errors.

La generación automática de ítems (GAI) es una metodología de elaboración de tests consistente en la creación de bancos de ítems a través de algoritmos computacionales. El primer objetivo de la GAI es crear ítems válidos para un test. En segundo lugar, se busca que los ítems tengan propiedades psicométricas predichas de antemano, sin necesidad de calibrarlos a partir de respuestas reales (Bejar, 1993, 1996; Revuelta y Ponsoda, 1998a). La GAI constituye un punto de encuentro entre distintas disciplinas: su correcta aplicación depende de la elaboración de teorías sobre la forma en que los sujetos resuelven los ítems. Esta teoría debe traducirse en un modelo psicométrico que relacione las propiedades de los ítems con la forma en que los sujetos los resuelven. Finalmente, es necesario crear un método informático que realmente elabore los ítems.

Desde un punto de vista práctico la GAI está motivada por la gran demanda de ítems que suponen los Tests Adaptativos Informatizados (TAIs) (Wainer, 1990). Para elaborar un TAI es necesario contar con un amplio banco de ítems, el cual debe renovarse periódicamente para evitar la difusión de su contenido (Revuelta y Ponsoda, 1996, 1998b). Este proceso de creación y renovación continua del banco de ítems resulta sumamente costoso por los recursos humanos y materiales que se necesitan para crear y calibrar los ítems, por lo que un sistema automatizado contribuiría a reducir significativamente los costes.

Correspondencia: Javier Revuelta
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
E-mail: Javier.Revuelta@uam.es

Una alternativa para aplicar la GAI cuando no se posee una teoría exhaustiva del modo en que los sujetos abordan la tarea, es mediante el proceso de ítems isomorfos. Para aplicar este método es necesario contar con una muestra de ítems calibrados a partir de respuestas reales. Cada uno de estos ítems se utiliza para crear ítems *isomorfos*, es decir ítems de aspecto similar (pero no idéntico) al original. Los isomorfos se crean modificando aquellas características del contenido de los ítems que se supone no afectan a sus propiedades psicométricas. A cada uno de los ítems isomorfos se le asignan los mismos parámetros que al ítem inicial modificados según se comenta mas adelante. Por ejemplo, supongamos un ítem de cálculo en el que al sujeto se le plantee la siguiente pregunta:

Indique el resultado de la expresión

$$\frac{\partial}{\partial x}\frac{x^2}{2} = ?$$

Puede suponerse que los elementos responsables de la dificultad de este ítem son únicamente las operaciones *derivada* y *cociente*, por lo que pueden elaborarse ítems isomorfos variando la magnitud de los valores numéricos que aparecen en el ítem. De acuerdo con este planteamiento la dificultad de todos estos isomorfos sería la misma que la del ítem original.

El proceso de creación de isomorfos introduce una nueva fuente de error en los parámetros de los ítems debido a que la manipulación del contenido puede afectar a las propiedades del ítem de una forma no controlada. Continuando con el ejemplo anterior, cabe la posibilidad de que la dificultad de los isomorfos dependa en

parte de cual sea la magnitud de los valores numéricos del exponente de *x* y el denominador.

El objetivo de este artículo es obtener información acerca de como el error en los parámetros de los isomorfos afecta a la fiabilidad de las puntuaciones en un TAI aplicado a partir de un banco de ítems que incluya isomorfos. A continuación se comenta el modelo psicométrico empleado para asignar parámetros a los isomorfos a partir de los del ítem original. En segundo lugar se describe estudio de Monte Carlo en el que se manipula sistemáticamente el error cometido en la asignación de parámetros y se obtienen estimaciones del sesgo y la eficiencia del estimador de la puntuación en el test.

Metodología de las Funciones Esperadas de Respuesta

La metodología de las Funciones Esperadas de Respuesta (FER) ha sido propuesta recientemente por Mislevy, Wingersky y Sheehan (1994) con el objetivo de establecer un método para incorporar la incertidumbre acerca de los parámetros de los ítems en la calibración del test. Esta metodología también ha sido aplicada al problema de hacer uso de información previa en el proceso de estimación de dichos parámetros (Mislevy, 1988; Mislevy, Sheehan y Wingersky, 1993).

Supongamos que f (Y | θ, ε) es la función psicométrica utilizada para calibrar los ítems, donde *Y* es una variable aleatoria que indica la respuesta del sujeto, θ es el nivel de rasgo del sujeto y ε es el vector de parámetros del ítem. La forma de f (Y | θ, ε) es arbitraria en la aplicación de la FER. En el caso más sencillo puede ser uno de los conocidos modelos logísticos de 1, 2 o 3 parámetros (Hambleton y Swaminathan, 1985).

Tradicionalmente para elaborar un test se obtiene el estimador de ε, denominado ε*, y se sustituye en f (Y | θ, ε) para realizar inferencias sobre θ a partir de las respuestas de cada sujeto. El inconveniente de este procedimiento es que no tiene en cuenta la incertidumbre asociada al proceso de estimación de ε. La consecuencia es que puede producirse una sobrestimación de la información disponible sobre θ en la aplicación del test (Mislevy et al., 1993).

La incertidumbre acerca del valor real de ε se traduce en que no es posible asignarle un valor puntual, sino una distribución de valores denominada g (ε | ε* , Σ). Los autores mencionados consideran la incertidumbre sobre ε debida al proceso de estimación. Entonces, utilizando una aproximación normal al estimador de ε (Tanner, 1993), la función g (ε | ε* , Σ) sería la distribución normal multivariada centrada en ε* y Σ sería la matriz de covarianza del estimador.

La anterior definición implica que ε es una variable aleatoria, por lo que la distribución conjunta de Y y ε es:

$$h(Y, \varepsilon \mid \theta) = f(Y \mid \theta, \varepsilon)\, g(\varepsilon \mid \varepsilon^*, \Sigma)$$

El vector ε es una variable no observada, por lo que se define la probabilidad marginal de *Y* como la función de respuesta, en sustitución de f (Y | θ, ε):

$$h(Y \mid \theta) = \int f(Y \mid \theta, \varepsilon)\, g(\varepsilon \mid \varepsilon^*, \Sigma)\, d\theta$$

La función h (Y | θ) es la Función Esperada de Respuesta. Como puede apreciarse h (Y | θ) no es mas que la función psicométrica promediada sobre la distribución de incertidumbre de ε*. Operativamente no suele utilizarse la FER según su definición, sino que se aproxima utilizando el modelo paramétrico que en f (Y | θ, ε). Por ejemplo, supongamos que f (Y | θ, ε) es el modelo logístico de 3 parámetros (3pl). Se obtiene un estimador máximo verosímil de ε, denominado ε*, y la correspondiente matriz de covarianza de este estimador S. A continuación se calcula h (Y | θ) y se busca la curva correspondiente al 3pl que sea la mejor aproximación a h (Y | θ). Los parámetros de esta curva se denominan ε**, y pueden utilizarse en f (Y | θ, ε) en sustitución de ε*. Típicamente esto produce una curva característica del ítem más plana que la que se obtiene utilizando ε* en la función de respuesta, como efecto de la incertidumbre acerca de ε.

En este trabajo se utiliza la FER para incorporar en la función de respuesta la incertidumbre sobre ε procedente del proceso de creación de isomorfos, no de su estimación como en la discusión precedente. Si se utiliza un ítem para crear varios isomorfos puede definirse la función g (ε | ε* , Σ) como la función de densidad de los parámetros de estos isomorfos, centrada en los parámetros del ítem original ε*, y con una matriz de error Σ procedente de la manipulación del contenido del ítem para crear sus isomorfos.

Idealmente, pueden crearse varios isomorfos a partir de un ítem, aplicarse a una muestra de sujetos y obtener una estimación de g (ε | ε* , Σ). A partir de esta estimación se obtiene ε** y se utiliza como parámetros de estos isomorfos y de otros nuevos que pudieran crearse para el mismo ítem. Sin embargo estos datos no están disponibles en la presente investigación. Este artículo se centra en el estudio de como la magnitud de los errores en Σ afecta a la fiabilidad de las puntuaciones en un TAI aplicado a partir de los ítems isomorfos. Como función g (ε | ε* , Σ) se escoge a priori la distribución normal multivariada

Se utilizó un banco de ítems de conocimientos gramaticales, compuesto por 363 elementos calibrados bajo el modelo 3pl. Estos ítems se utilizaron para crear isomorfos bajo distintas condiciones de error. El proceso de creación de isomorfos afecta únicamente a sus aspectos psicométricos, no al contenido de los ítems. Consiste en obtener los parámetros de los isomorfos bajo distintas magnitudes de los errores en la matriz Σ.

El banco se utilizó para aplicar TAIs mediante simulación, y estimar la fiabilidad de las puntuaciones en el mismo. Los TAIs se aplicaron en condiciones realistas en cuanto a su longitud, restricciones de contenido en la selección de los ítems (Stocking y Swanson, 1993) y control de la exposición.

Estudio de simulación

El objetivo es comprobar el efecto de la magnitud de los errores en la fiabilidad de las puntuaciones. Debido a la falta de información empírica sobre Σ se decidió utilizar la matriz de covarianza de los parámetros estimados en los 363 ítems como base del estudio. Esta matriz proporciona una estimación realista de la magnitud relativa de las varianzas y covarianzas entre los distintos parámetros. Como el banco está calibrado de acuerdo con el 3pl la composición del vector ε es ε = (a, b, c), donde *a*, *b* y *c* son los parámetros de discriminación, dificultad y adivinación del ítem. La matriz de covarianza es:

$$\Sigma = \begin{bmatrix} 0.2376 & & \\ 0.2208 & 1.2749 & \\ 0.0170 & 0.0252 & 0.0083 \end{bmatrix}$$

Condiciones

Se manipularon dos variables independientes: la primera es la magnitud de los errores, con cuatro niveles. Se multiplicó la matriz Σ por una constate *C* que toma para cada condición los valores: 0.5, 1, 1.5 y 2. La matriz resultante se utilizó como matriz de error en la función $g(\varepsilon \mid \varepsilon^*, S)$. La segunda variable independiente es el porcentaje de ítems con isomorfos en el banco. Esta variable tiene los niveles 25, 50 75 y 100. Se cruzaron los 4 niveles de las dos variables independientes, obteniéndose 16 condiciones. Además se incluyó un grupo control sin isomorfos para obtener una estimación en el mejor caso y comparar los resultados de los demás grupos.

Procedimiento

Para cada ítem con isomorfos el valor de ε^* es el estimador de sus parámetros, es decir es específico del ítem. En cambio, el valor de Σ es común en cada condición para todos los ítems del banco que originan isomorfos. En cada una de las 16 condiciones con isomorfos se utilizó la metodología FER para calcular los parámetros atenuados ε^{**} de los isomorfos. Estos parámetros atenuados se calcularon mediante el programa EXPRESFN (Mislevy et al., 1994). A continuación se aplicó en las 17 condiciones un TAI de 28 ítems a un grupo de 5500 sujetos simulados. Estos sujetos se dividían en 11 grupos de 500 sujetos, con puntuación verdadera igual a 10, 15, 20, 25,… , 55, y 59. La habilidad verdadera y estimada se proporciona en la escala de las puntuaciones en el test, no en la escala θ.

Los ítems en el banco se clasificaban en ítems normales e ítems con isomorfos. Cada vez que se selecciona un ítem con isomorfos para ser administrado, se genera un valor aleatorio de la distribución $g(\varepsilon \mid \varepsilon^*, \Sigma)$, que se considera como el vector verdadero de parámetros del ítem y se utiliza en la obtención por simulación de la respuesta *Y*. Por el contrario, en los ítems normales la respuesta se genera a partir de los parámetros en ε^*. En la estimación de habilidad se utiliza el vector de parámetros estimado para el ítem, es decir los parámetros atenuados ε^{**} para los isomorfos y ε^* para los ítems originales del banco.

Variables y análisis de datos

En cada grupo se calculó la puntuación media y la función de información de dicha puntuación. Se comparó la información en el grupo sin isomorfos con la información en los demás grupos para estimar la pérdida producida por cada nivel de error. Se calculó la proporción de sujetos en cada grupo a los que se administraban 10 o más isomorfos para comprobar si estos items se utilizan realmente en el test adaptativo. Por último, se estimó la recta de regresión lineal de la fiabilidad en el test sobre la constante *C*, por separado para cada condición de la variable *porcentaje de isomorfos*. El objetivo es comparar el modo en que varía la fiabilidad en función de *C* en cada uno de estos grupos.

Resultados

La diferencia entre la puntuación verdadera y la estimada es mínima en todas las condiciones. Es decir, el error en el proceso de creación de isomorfos no produjo sesgo en los estimadores de habilidad. La tabla 1 muestra la puntuación verdadera y la estimada en la condición con un 25% de isomorfos, el orden de magnitud de la diferencia entre ambas es similar en el resto de condiciones.

Con respecto a la eficiencia de los estimadores, se encontró una relación muy clara con las dos variables independientes. La información del test disminuye a medida que aumenta el error en los isomorfos y también con el porcentaje de estos en el banco. Además se encontró un efecto de interacción. En cada grupo de la variable *porcentaje de isomorfos* la información disminuye al aumentar el error en Σ, sin embargo, este efecto es mas pronunciado a medida que aumenta el porcentaje de isomorfos en el banco.

En la condición con un 25% de ítems con isomorfos la información del test es aproximadamente el 80% de la que proporciona en la condición sin isomorfos (ver tabla 1). Dicho de otra manera, si se incluye este 25% de isomorfos es necesario alargar la longitud del test un 20% para compensar la pérdida de precisión introducida por la nueva fuente de error en los parámetros de los ítems. Con un 50% y un 75% de ítems con isomorfos la información se reduce al 40% y 60% respectivamente. Por último, en la condición con un 100% de isomorfos el test solamente proporcio-

Tabla 1
Resultados en las condiciones de control y con un 25% de isomorfos

V	M_0	$M_{0.5}$	M_1	$M_{1.5}$	M_2	I_0	$IR_{0.5}$	IR_1	$IR_{1.5}$	IR_2
59	59.45	59.37	59.30	59.21	59.12	1.32	1.02	1.07	1.12	1.07
55	55.01	54.97	54.95	54.95	55.00	0.33	0.84	0.78	0.72	0.73
50	50.11	50.13	50.14	50.29	50.15	0.17	0.89	0.90	0.83	0.85
45	45.10	45.20	45.07	45.36	45.10	0.14	0.80	0.76	0.77	0.83
40	40.22	40.21	40.35	40.24	40.35	0.11	0.87	0.88	0.80	0.82
35	34.90	34.87	34.82	34.82	34.69	0.10	0.81	0.77	0.81	0.84
30	30.02	29.85	29.81	29.89	29.89	0.11	0.83	0.80	0.77	0.75
25	25.07	25.09	25.14	25.19	24.99	0.13	0.89	0.76	0.70	0.83
20	20.16	20.24	20.28	20.30	20.30	0.13	0.94	0.88	0.83	0.81
15	15.08	15.01	14.97	15.01	15.03	0.15	1.03	0.98	0.93	0.89
10	13.30	10.28	10.32	10.26	10.30	0.42	1.03	0.96	1.01	0.94

V: Puntuación verdadera.
M_0: Media de las puntuaciones estimadas en el grupo de control.
$M_{0.5}$ a M_2: Media en las condiciones C=0.5 a C=2. Grupo con un 25% de isomorfos.
I_0: Información media del test en el grupo de control.
$IR_{0.5}$ a IR_2: Información relativa del test en la las condiciones C=0.5, C=1, C=1.5 y C=2 en comparación con la condición de control. Grupo con un 25% de isomorfos.

na un 18% de la información obtenida sin isomorfos (ver tabla 2). Para interpretar adecuadamente estos resultados debe tenerse en cuenta que un error de 2Σ en los isomorfos es muy elevado y no es esperable encontrarlo en la práctica. Esta condición se incluyó para obtener una estimación en el peor caso, aunque raramente se encuentren estos errores en la realidad.

En los cuatro grupos de la variable *porcentaje de isomorfos*, la pendiente estimada de la regresión fue negativa, lo que indica que la fiabilidad disminuye con el aumento en el error. Además la magnitud de la pendiente depende del porcentaje de isomorfos aumenta con el porcentaje de isomorfos. La constante de la regresión tomó en los cuatro grupos un valor próximo a 0.93, que es la fiabilidad del test en el caso de que la constante *C* sea 0 y no haya isomorfos (ver tabla 3). Estos resultados que indican que la degradación de la fiabilidad al aumentar los errores de los isomorfos es mayor a medida que aumenta el número de isomorfos en el banco.

Por último, el hecho de que aumente el número de isomorfos en el banco de ítems no significa que aumente el número de estos realmente administrados a los sujetos. Para comprobar este punto se calculó el porcentaje de sujetos que reciben mas de 10 isomorfos en el test. Estos porcentajes son los siguientes para los cuatro grupos de la variable *porcentaje de isomorfos*: 10%, 85%, 100% y 100%. Es decir, los ítems isomorfos realmente aparecieron con elevada frecuencia en los tests administrados.

Conclusiones

Esta investigación muestra que el proceso de creación de isomorfos puede ser una alternativa viable en el desarrollo de tests, en términos de las propiedades psicométricas del estimador de habilidad. Como cabe esperar, los errores introducidos en los parámetros de los ítems producen una disminución en la fiabilidad del test. Sin embargo esta disminución es pequeña si el proceso de asignación de los parámetros de los isomorfos es mínimamente preciso, y puede compensarse aumentando la longitud del test. Debe tenerse en cuenta que de las condiciones probadas la mas realista es aquella en que la matriz de errores en los isomorfos es la mitad de la matriz en el banco total de ítems. En aplicaciones prácticas cabe esperar matrices como esta o incluso menores, y en estas condiciones la pérdida de fiabilidad es pequeña.

La metodología FER resulta ser adecuada para la aplicación de la GAI en el caso de que se posea un conocimiento incompleto del modo en que se resuelven los ítems. En esta investigación la FER únicamente se aplica para incorporar la incertidumbre inherente al proceso de elaboración de isomorfos. No obstante, esta metodología puede también aplicarse de forma rutinaria en cualquier test, se base o no en la GAI, para incorporar el error de estimación de los parámetros de los ítems.

Para continuar esta investigación resulta necesario utilizar condiciones mas realistas. Por ejemplo, resulta necesario elaborar y

Tabla 2
Información relativa en función del número de isomorfos y de la constante C

	50% isomorfos				75% isomorfos				100% isomorfos			
V	$IR_{0.5}$	IR_1	$IR_{1.5}$	IR_2	$IR_{0.5}$	IR_1	$IR_{1.5}$	IR_2	$IR_{0.5}$	IR_1	$IR_{1.5}$	IR_2
50	1.15	1.12	1.18	1.38	1.18	1.35	1.35	1.02	1.24	1.46	1.42	0.12
55	0.69	0.69	0.62	0.62	0.66	0.47	0.41	0.36	0.55	0.45	0.28	0.18
50	0.77	0.64	0.61	0.54	0.60	0.47	0.39	0.34	0.48	0.32	0.25	0.18
45	0.66	0.63	0.58	0.62	0.50	0.40	0.35	0.32	0.37	0.30	0.21	0.17
40	0.80	0.59	0.57	0.60	0.54	0.43	0.36	0.34	0.48	0.28	0.22	0.17
35	0.76	0.64	0.58	0.52	0.55	0.44	0.37	0.31	0.48	0.31	0.23	0.17
30	0.73	0.64	0.56	0.52	0.53	0.41	0.35	0.27	0.44	0.31	0.19	0.18
25	0.65	0.59	0.58	0.50	0.48	0.39	0.33	0.33	0.45	0.29	0.22	0.16
20	0.78	0.65	0.63	0.55	0.66	0.48	0.37	0.35	0.54	0.39	0.23	0.24
15	0.79	0.62	0.59	0.58	0.73	0.62	0.60	0.52	0.65	0.54	0.41	0.35
10	0.98	1.51	0.76	0.85	1.01	0.79	0.75	0.69	1.03	0.80	0.73	0.76

V : Puntuación verdadera.
$IR_{0.5}$ a IR_2 : Información relativa del test en la las condiciones C=0.5, C=1, C=1.5 y C=2 en comparación con el grupo de control.

Tabla 3
Regresión lineal de la fiabilidad en el test sobre el valor de la constante C. Resultados por separado para cada una de las condiciones de 25%, 50%, 70% y 100% de isomorfos

Var	β	Sβ	t	P (t)	α	R^2
R_{25}	-0.0068	0.003	-2.65	0.077	0.932	0.60
R_{50}	-0.0224	0.004	-6.30	0.008	0.930	0.90
R_{75}	-0.0542	0.007	-7.48	0.005	0.949	0.93
R_{100}	-0.1090	0.006	-17.48	0.000	0.990	0.99

β: Pendiente de la regresión.
Sβ: Error de estimación de b.
T: Estadístico de contraste.
P (t): Nivel crítico.
α: Constante de la regresión
R^2: Correlación múltiple ajustada.

calibrar los isomorfos a partir de su aplicación a sujetos reales, y obtener una estimación de la forma de la función $g(\varepsilon \mid \varepsilon^*, \Sigma)$ y el parámetro Σ. A continuación es posible obtener por simulación una estimación de la fiabilidad del TAI aplicado a partir de dichos ítems, repitiendo el proceso descrito en este artículo.

Nota

Esta investigación ha sido financiada en parte por el proyecto de la DGICYT PB 97-0049.

Referencias

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. En N. Frederiksen, R. L. Mislevy y I. I. Bejar (eds.). *Test theory for a new generation of tests*. Hillsdale: Erlbaum.

Bejar, I.I. (1996). Generative response modeling: leveraging the computer as a test delivery medium. *Research Report 96-13*. Educational Testing Service.

Hambleton, K. H. y Swaminathan H. S. (1985). *Item Response Theory. Principles and Applications*. Boston: Kluwer Nijhoff Pub.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement.* 12, 3, 281-296.

Mislevy, R. J., Sheehan, K. M. y Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement.* 30, 1, 55-78.

Mislevy, R. J., Wingersky, M. y Sheehan, K. M. (1994). Dealing with uncertainty about item parameters: Expected Response Functions. *Research Report 99-28-ONR*. Educational Testing Service.

Revuelta, J. y Ponsoda, V. (1996). Métodos sencillos para el control de las tasas de exposición en tests adaptativos informatizados. *Psicológica*, 17, 1, 161-172.

Revuelta, J. y Ponsoda, V. (1998a). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement.* 35, 4, 311-328

Revuelta, J. y Ponsoda, V. (1998b). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema*. 10, 3, 709-716.

Stocking, M. y Swanson, L. 1993. A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.

Tanner, M. A. (1993). *Tools for statistical inference*. New York: Springer-Verlag.

Wainer, H. (1990). *Computerized adaptive testing: a primer*. Hillsdale: LEA.

Aceptado el 19 de julio de 1999