

EFFECTOS DE LA VIOLACIÓN DE LOS SUPUESTOS DEL MODELO DE RASCH SOBRE LA ROBUSTEZ DE SUS ESTIMACIONES

M^a Isabel Barbero García* y Pedro Prieto Marañón**

* UNED, **Universidad de La Laguna

En el presente trabajo se pretende evaluar la robustez de las estimaciones de b y θ mediante el modelo de Rasch cuando existen aciertos al azar ($c \neq 0$), y los índices de discriminación de los ítems no son constantes ($a \neq K$). Para ello se ha realizado una réplica del trabajo de Muñíz, Rogers y Swaminathan (1989), en el que se introducen tres cambios: El programa de generación de datos —realizado ad hoc para este fin—, el programa de estimación de parámetros —BILOG—, así como el programa de análisis de ajuste de las curvas características, elaborado por los autores. La finalidad última que se persigue, además de la señalada en primer lugar, es la de evaluar la validez de los programas desarrollados de cara a su posterior aplicación en diversos estudios dentro de los modelos de rasgos latentes.

Effects of the violation of the rasch model assumptions on the robustness of estimations. Following the work of Muñíz, Rogers y Swaminathan (1989), in the present paper we test the robustness of parameters b y θ estimations in the Rasch model when some assumptions are violated ($c \neq 0$ y $a \neq K$). However, three changes are introduced now: Programs for examinee response data generating and checking models goodness of fit —elaborated by authors— and estimation parameters through BILOG program. What is intended, at the end, is checking our new programs for further applications in some researchs in latent trait models.

El procedimiento general habitualmente seguido en los estudios de simulación llevados a cabo dentro del campo de la TRI puede dividirse en tres fases: Una primera, en la que una vez decidido los valores de los parámetros a manipular, se generan de forma aleatoria los datos de los distintos bloques que conformarán las diferentes condiciones de la investigación. En una segunda fase, y a partir de estos bloques de datos, se estiman los valores de los parámetros

de los modelos elegidos mediante alguno de los programas que a tal efecto existen en el mercado (BICAL, BILOG, LOGIST, MICROCAT, etc...). Por último se evalúa la precisión de tales estimaciones, bien a través de correlaciones entre los valores simulados y los estimados, bien a través de índices de bondad de ajuste.

Este proceso de investigación, en el que se emplean, al menos, tres diferentes programas en cada una de las tres fases, requiere una dependencia del investigador, en tanto en cuanto, una vez finalizada cada una de estas tres fases, debe preparar los datos y programas para proceder a la siguiente fase. Para el investigador sería

Correspondencia: M^a Isabel Barbero García
Facultad de Psicología
UNED. Ciudad Universitaria
Madrid. Spain

un gran ahorro de tiempo y trabajo, el disponer de un único programa encargado de llevar a cabo de forma automática e ininterrumpida todo el proceso.

Otra consideración importante es la referida a la *desconfianza* que ante los programas de simulación de datos debe tener siempre en cuenta el investigador. A menos que el simulador a emplear nos ofrezca las debidas garantías, el investigador debería previamente a su estudio verificar la bondad de ajuste del mismo, puesto que no sería la primera ocasión en la que se demostrara que los datos generados por el mismo no cumplen los requisitos exigidos por la investigación.

Estando interesados en llevar a cabo algunos estudios dentro del marco de la TRI mediante procedimientos de simulación, y atendiendo a lo anteriormente expuesto, hemos pretendido diseñar un programa global (GENESTE) encargado de llevar a cabo las tres fases anteriormente descritas (GEN = generación, EST = estimación, E = evaluación). Para ello, hemos desarrollado para la primera y tercera de las fases, nuestro propio simulador de datos (generador de congruencia lineal multiplicativo puro, con módulo $2^{31} - 1$ y multiplicador 1343714438) y un programa de análisis de la precisión de las estimaciones (índices de ajuste, residuos, chi cuadrado). Para el segundo paso, de estimación de parámetros, hemos recurrido al programa BILOG (Mislevy y Bock, 1984).

Los tres programas han sido acoplados como módulos dentro de un programa general, desarrollado en TURBO-C, que controla todo el proceso. Mediante el mismo, sólo es necesario definir inicialmente los parámetros a manipular, forma de las distribuciones, modelos logísticos a emplear, nombre de ficheros de resultados, etc... Una vez introducidos estos datos el proceso de generación de datos simulados, estimación de parámetros y cálculo de

los índices de precisión se desarrolla ininterrumpidamente hasta su conclusión.

Con el fin de poner a prueba los programas elaborados hemos replicado, en el trabajo que hoy exponemos, el estudio llevado a cabo por Muñíz, Rogers y Swaminathan en 1989, en el que mediante la utilización de datos simulados se pretende evaluar la robustez de las estimaciones de b y Θ mediante el modelo de Rasch, cuando se dan aciertos al azar ($c \neq 0$), y los índices de discriminación de los ítems no son constantes ($a \neq K$), es decir, ante la violación de dos de los supuestos básicos de dicho modelo.

La idea de estudiar las influencias de estas violaciones del modelo sobre las estimaciones del mismo proviene de la relativa imprecisión en la estimación de los valores de a y c con muestras pequeñas (50 ítems; 500-1000 sujetos) según se desprende de los trabajos de autores como Kolen (1981), Ree (1979), Thissen y Wainer (1982), así como de las críticas de Wright (1977) al uso de modelos logísticos de dos y tres parámetros. De ahí surge la cuestión de si en estos casos en que se dispone de muestras pequeñas sería aconsejable utilizar el modelo de Rasch en el que sólo es necesario estimar el parámetro b para cada uno de los ítems, como aconseja Lord (1983).

La sencillez del modelo de Rasch hace que sea atractivo para los diseñadores de instrumentos de medición lo cuál conlleva su uso extendido, aplicándose incluso en casos de ítems de elección múltiple en los que las evidencias y el simple sentido común muestran que ni el índice de discriminación es constante para todos los ítems que integran la prueba, ni el valor del parámetro c de aciertos al azar es cero.

Simulación de los datos

Para llevar a cabo el estudio se generaron 12 bloques de datos de 50 ítems x 500 sujetos cada uno, en los cuales se

variaban los valores de los parámetros a y c . En seis de estos bloques el valor de « a » fue constante e igual a 1, mientras que en los seis restantes el valor de « a » se distribuía uniformemente entre valores de 0.5 y 2 en cada uno de los bloques.

$N(0,1)$, los mismos para los 12 bloques, y los índices de dificultad « b » se generaron con distribución uniforme entre -2 y +2 para los mismos 12 bloques.

Una vez establecidas estas condiciones (de forma similar a las planteadas por

Tabla 1
Precisión de las estimaciones

Blo.	c	a	r $\theta\theta'$			rbb'			IA(Θ)			Res.<Abs.(2)		
			1p	2p	3p	1p	2p	3p	1p	2p	3p	1p	2p	3p
1	0	1	973	973	973	998	995	996	253	250	253	947	983	985
2	0-.5	1	929	939	943	903	833	967	372	349	349	543	965	985
3	.5	1	872	897	902	972	923	982	486	437	448	361	963	982
4	.33	1	909	922	928	987	982	984	414	388	377	400	910	923
5	.25	1	934	936	948	991	976	992	357	354	322	588	960	990
6	.20	1	947	949	956	991	979	992	321	318	300	578	968	977
7	0	.5-2	975	978	977	987	997	997	263	259	256	875	980	988
8	0-.5	.5-2	946	957	960	927	928	982	321	304	284	598	943	982
9	.5	.5-2	878	918	925	968	894	983	474	392	402	421	925	982
10	.33	.5-2	920	935	950	978	963	983	391	359	323	474	932	983
11	.25	.5-2	937	946	961	973	973	993	349	333	296	583	933	981
12	.20	.5-2	937	945	963	982	958	991	352	343	302	660	925	971

En cada uno de estos dos grupos de seis bloques, los valores de « c » tomaron los siguientes valores: 0, .20, .25, .33 y .50 para todos los items y distribución uniforme entre 0 y .5.

Por su parte los valores Θ de los sujetos se generaron con distribución normal

Muñiz et al.) se introdujeron dichos datos en el programa diseñado, el cual, como ya hemos puesto de relieve, se encargó de realizar todo el proceso de manera ininterrumpida. Es decir, generación de matrices de datos simuladas, estimación de los parámetros de los modelos logísti-

Tabla 2
Intercorrelaciones entre los valores simulados y estimados

Bloq.	c	a	r $\theta\theta'$			rbb'		
			r12	r13	r23	r12	r13	r23
1	0	1	9996	9995	9997	9974	9977	9997
2	0-.5	1	9901	9812	9286	9491	9343	9034
3	.5	1	9825	9698	9939	9469	9555	9132
4	.33	1	9904	9740	9888	9644	9714	9687
5	.25	1	9916	9833	9901	9586	9838	9821
6	.20	1	9940	9864	9920	9659	9884	9852
7	0	.5-2	9973	9969	9996	9865	9857	9996
8	0-.5	.5-2	9914	9839	9928	9607	9596	9565
9	.5	.5-2	9774	9490	9883	8522	9616	9684
10	.33	.5-2	9879	9687	9858	9332	9698	9633
11	.25	.5-2	9898	9734	9847	9339	9678	9739
12	.20	.5-2	9888	9762	9863	9363	9808	9722

cos de 1, 2 y 3 parámetros y evaluación de la precisión de las estimaciones mediante las correlaciones entre los valores simulados de los parámetros y los estimados, índices de ajuste y bondad de ajuste de las curvas características (residuos estandarizados, chi-cuadrado).

Resultados y conclusiones

Los resultados obtenidos en este estudio coinciden ampliamente, en líneas

generales, con los encontrados en el trabajo de Muñiz et al. (1989).

Respecto a la influencia de los valores del parámetro «c» sobre la precisión de las estimaciones del modelo de Rasch se observa que, cuando esta precisión se mide a través de las correlaciones entre los valores simulados-estimados decae ligeramente al incrementarse los aciertos al azar. También ocurre lo mismo con el modelo de tres parámetros debido, tal vez, a la imprecisión contaminadora de

Tabla 3
Intercorrelaciones entre los valores simulados y estimados

Bloq.	c	a	rcc'		raa'	
			3p	2p	3p	2p
2	0-.5	1	7922			
7	0	.5-2		9091	8882	
12	.20	.5-2		3494	8062	
11	.25	.5-2		2866	7839	
10	.33	.5-2		3517	7631	
9	.5	.5-2		2421	7687	
8	0-.5	.5-2	7283	3751	7205	

las estimaciones de c , con correlaciones entre los valores simulados y estimados de 0.79 (0.39 Muñiz) para $a = 1$ y de 0.72 (0.60 Muñiz) para $a = 0.5 - 2$ (ver tabla 3).

En cualquier caso la precisión de las estimaciones se mantiene similar en los modelos de 1 y 3 parámetros.

Cuando se analiza el ajuste de los modelos a través del análisis de residuales, el ajuste general del modelo de tres parámetros es sensiblemente mejor que el del modelo de Rasch (la diferencia es sensiblemente superior a la del estudio de Muñiz).

Podemos coincidir con van der Vijver (1986) que las violaciones del supuesto de $c = 0$ parecen afectar más al ajuste del modelo como discrepancia entre los valores simulados-estimados que a las correlaciones entre dichos valores.

Cuando se analiza el índice de ajuste IA (Θ), se observa también una ligera superioridad del mismo para el modelo de 3 parámetros a medida que se incrementa el valor de c .

En cuanto a la cuantía de los desajustes (evaluados a través de los residuos) en función de los valores de Θ , se observa que cuando $c \neq 0$ el ajuste es peor para los valores bajos de Θ , mejorando para valores medios y altos en el caso del modelo de Rasch.

De estos datos podemos concluir la aparente robustez del modelo de Rasch frente a violaciones de $c = 0$ en cuanto a la precisión de las estimaciones entendida como correlación simulado-estimado, robustez que se ve mermada si entendemos tal precisión como discrepancia simulado-estimado.

Respecto a la robustez del modelo de Rasch en el caso de no igualdad de los índices de discriminación al compararlo con la condición $a = K$, de los cuatro índices de precisión, se mantienen parejos los índices de correlación, empeorando ligeramente en cuanto a IA(θ) y

sensiblemente en lo que a la proporción de residuos. Por otra parte, la ganancia de precisión proporcionada por los modelos de dos y tres parámetros, que serían los adecuados en estas condiciones, es escasa en general, aunque algo mayor considerada a nivel de residuos. A nivel de índices de precisión correlacionales se puede mantener la afirmación de van der Vijver (1986) y Muñiz et al. (1989) en cuanto a que la presencia de índices de discriminación variables no parecen afectar significativamente a la precisión de las estimaciones del modelo de Rasch, no pudiendo mantener tal afirmación en lo referente a los índices de precisión referidos a la discrepancia simulado-estimado (IA (θ) y residuos).

En cualquier caso, a través de los datos obtenidos puede comprobarse cómo las violaciones del supuesto de discriminación constante parecen afectar a la precisión de las estimaciones en menor medida que las violaciones del supuesto de que $c = 0$.

Por último, la interacción de las violaciones de c y a no aportan ninguna novedad relevante. No parece que al darse ambos tipos de violaciones conjuntamente se produzca un efecto multiplicativo sobre la imprecisión de las estimaciones del modelo de Rasch. De nuevo, los índices de precisión se mantienen similares a niveles correlacionales, empeorando en lo referente a la proporción de residuos estandarizados, siendo de nuevo mejor el ajuste en el modelo de tres parámetros.

Un caso bastante frecuente sería aquél en el que los índices de discriminación son variables y los valores de c son distintos de cero (entre 0 y 0,5). En este caso de total violación de los supuestos del modelo de Rasch, éste se desenvuelve bastante bien cuando se considera la precisión del modelo a nivel correlacional. En tal caso, las ganancias de precisión

generadas por el modelo de tres parámetros (teóricamente más indicado) frente al de Rasch son escasamente de 0,0144 y 0,0543 para $r_{\theta\theta'}$ y $r_{bb'}$ respectivamente. Sin embargo a nivel de proporción de residuos esta ganancia sería de 0,38.

En todo caso, las altas correlaciones entre las estimaciones de los mismos parámetros b y θ por los tres modelos, puede sugerirnos las escasas ventajas a nivel correlacional de los modelos de dos y tres parámetros frente al de Rasch. Como ya se ha señalado, los datos sugieren una vez más que si el interés del usuario se centra en el escalamiento de sujetos o items, la elección de un modelo u otro va a tener una incidencia mínima en el ordenamiento de los mismos.

Finalmente, podemos comprobar la imprecisión de las estimaciones de a y c . Las estimaciones de a (especialmente en el modelo de dos parámetros, y algo menos en el de tres) resultan bastante imprecisas a medida que c se aleja del supuesto $c=0$. Podría decirse que las estimaciones de a en el modelo de dos parámetros son muy sensibles a la presencia de aciertos al azar. A su vez, las estimaciones de c son así mismo ligeramente imprecisas (con correlaciones de 0,72 y 0,79), confirmándose lo apuntado con anterioridad por diversos autores.

Como conclusiones de este breve estudio, podemos mantener, al igual que

Muñíz et al., que las estimaciones del modelo de Rasch, y por lo tanto su ajuste, se ven poco afectadas por la presencia de aciertos al azar y menos aún por la variabilidad de los índices de discriminación, cuando tomamos en cuenta los índices de precisión correlacionales. Así mismo, la ganancia de precisión proporcionada por los modelos de dos y tres parámetros es escasa en el mejor de los casos. Esta afirmación se ve un tanto cuestionada cuando se utilizan como índices de precisión los residuos estandarizados.

En cualquier caso, y como ya se ha reseñado, la elección del modelo de Rasch en tales circunstancias de violación de los supuestos, no parece influir sobre la precisión del escalamiento de estímulos y/o sujetos.

A su vez, en el caso del uso de modelos más complejos, la ganancia de precisión que estos pueden ofrecer se ve mermada por la imprecisión de las estimaciones de a y c .

Como conclusión general podemos considerar que los programas que hemos desarrollado han respondido satisfactoriamente al propósito para el que fueron diseñados, pudiendo disponer, por lo tanto, de un interesante instrumento para el desarrollo de futuras investigaciones.

Referencias

- Kolen, J. (1981). Traditional equating methodology. *Educational Measurement*, Vol. 7, 4, 29-36.
- Lord, F.M. (1983). Small N justifies Rasch model. En R.J. Weiss (ed.): *New Horizons in Testing*, Nueva York, Academic Press.
- Mislevy, R.J. y Bock, R.D. (1984): *BILOG er. 2.2: Item analysis and test score with binary logistic models*. Mooresville, I.N., Scientific software.
- Muñíz, J.; Rogers, J. Swaminathan, H. (1989): Robustez de las estimaciones del modelo de Rasch en presencia de aciertos al azar y discriminación variable de los items. *Anuario de Psicología*, vol. 4, nº 3, 83-97.

- Ree, J.M. (1979). Estimating item characteristic curves. *Psychological Measurement*, 3, 371-385.
- Thissen, D.; Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47, 397-412.
- Van der Vijver, F.J. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement*, 10 (1), 45-57.
- Wright, B.D. (1977a). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.
- Wright, B.D. (1977b). Misunderstanding of the Rasch model. *Journal of Educational Measurement*, 14, 219-226.