Estimación de la magnitud del efecto en invarianza de medición

Sergio Dominguez-Lara[1], *César Merino-Soto*
Universidad de San Martín de Porres, Lima, Perú

RESUMEN

Este trabajo presenta un programa en MS Excel® para evaluar la magnitud del efecto (ES, por las siglas en inglés) en invarianza de medición de diferentes parámetros como las cargas factoriales, interceptos/thresholds y residuales, con base en estadísticos estandarizados ya conocidos. El funcionamiento del programa y la interpretación de las medidas de ES se ejemplificaron con datos empíricos.

Palabras clave: invarianza de medición; magnitud del efecto; programa.

RESUMO – Avaliação do tamanho do efeito na invariância de medida

Este artigo apresenta um programa no MS Excel® para avaliar o tamanho do efeito (ES, pela sigla em inglês) na invariância de medida de diferentes parâmetros dos itens, como cargas fatoriais, interceptos/thresholds e resíduos, com base em estatísticas padronizadas bem conhecidas. O funcionamento do programa e interpretação das medidas de ES é exemplificado com dados empíricos.

Palavras-chave: invariância de medida, tamanho do efeito, programa.

ABSTRACT – Estimation of effect size in measurement invariance

This paper presents a MS Excel® program for estimating effect size (ES) in the measurement invariance of different item parameters, such as the factor loadings, intercepts/thresholds and residuals, based in well-known standardized statistics. Examples with real data are provided for the functioning of the program and the interpretation of the ES measures.

Keywords: measurement invariance; effect size; computer program.

El análisis de *invarianza de medición* (IM) es un procedimiento valorado como una de las fuentes de evidencias para detectar la presencia de sesgo de medida. Para implementarlo desde el marco de la *teoría clásica de los tests*, se recomienda un enfoque de *evaluación gradual*, en que se aplica restricciones de igualdad de parámetros de manera acumulativa: invarianza configural, métrica, escalar, y estricta (Pendergast, von der Embse, Kilgus, & Eklund, 2017).

La evaluación de la invarianza inicia con examinar una prueba de bondad de ajuste (basada en el χ^2) y los índices de ajuste (e.g., CFI, RMSEA, etc.) derivados de la prueba χ^2 anterior. Cuando la variación de los índices de ajuste (e.g., diferencias en el CFI; Cheung & Rensvold, 2002) sugiere que no es plausible determinado grado de invarianza (e.g., métrica) se examinan las posibles *fuentes* de esta conclusión. Generalmente, se usan los *índices de modificación* (Sörbom, 1989), basados en el estadístico χ^2, para detectar qué ítems son no-invariantes, es decir, qué restricciones del modelo que parecen producir el decremento en el ajuste global (o desajuste). Por ejemplo, se puede evidenciar que la restricción de igualdad de cargas factoriales impuesta sobre un ítem en ambos grupos es la fuente de la ausencia de invarianza, lo que indica que dicho ítem representa de forma distinta el constructo en ambos grupos.

Por años, la base estadística de las decisiones sobre la hipótesis de invarianza ha pesado fundamentalmente sobre la prueba de significancia asociada a la χ^2, como el resto de los métodos de análisis estadísticos; pero actualmente, el reporte de estos resultados está subsumido en la tendencia moderna para comunicar efectivamente análisis cuantitativos (e.g., Appelbaum et al., 2018), que incluye la información de la *magnitud del efecto* (effect size; ES).

En el contexto de la IM, la ES cuantifica el grado en que los parámetros comparados son diferentes entre los grupos, ya que la prueba de significancia de la hipótesis nula (NHST), sobre la cual se interpretan los *índices de modificación*, no informa directamente sobre la magnitud

[1] Endereço para correspondência: Instituto de Investigación de Psicología, Universidad de San Martín de Porres. Avenida Tomás Marsano, 242, 5o. piso, Lima 34, Perú. E-*mail*: sdominguezl@usmp.pe; sdominguezmpcs@gmail.com

de la *no-invarianza* del parámetro, lo que puede llevar a concluir que la medida no es invariante, aún si la diferencia entre los parámetros del grupo de referencia (GR) y grupo focal (GF) es pequeña.

En este sentido, la magnitud de la diferencia entre los parámetros evaluados puede informar sobre la *significancia práctica* de cada uno. Un antecedente para este planteamiento ha sido expuesto en la comparación de medias latentes (Choi, Fan, & Hancock, 2009) con el estimador de diferencias de medias estandarizadas (*d*; Cohen, 1988) basado en criterios conocidos (.20, diferencia pequeña; .50, moderada; .80, grande), y más adelante se descubrió que esos fundamentos pueden aplicarse a los otros parámetros involucrados en la IM.

Para introducir una evaluación gradual de la ES en la IM, Pornprasertmanit (2014) dedujo que los parámetros implicados en la evaluación de la invarianza de medición (cargas factoriales, interceptos o *thresholds*, y residuales) pueden ser tratados como otros parámetros de naturaleza descriptiva, dado que tienen en común el rango fijo de sus valores. De este modo se deducen algunas propiedades del estadístico de interés, lo que es una práctica habitual en la literatura metodológica; por ejemplo, en la comparación de coeficientes α estos se tratan como varianzas (Feldt, 1980); algunos coeficientes de validez de contenido, como la V de Aiken, se consideran proporciones para deducir sus intervalos de confianza (Penfield & Giacobbi, 2004).

En este sentido, para la comparación de cargas factoriales, Pornprasertmanit (2014) propuso usar el estimador *q* (Cohen, 1988), aplicado originalmente para comparar correlaciones (.10, diferencia pequeña; .30, moderada; .50, grande); en este contexto se aplica bajo el supuesto de que la *carga factorial* es una medida de asociación entre el ítem y el factor, es decir, entre la *puntuación observada* y la *variable latente*; además, podría analizarse como el grado influencia del factor sobre el ítem, lo que haría viable su comparación al transformar ese indicador en una puntuación Z (DeLong & Elbeck, 2018).

Por otro lado, para la comparación de interceptos y *thresholds* se calcula una diferencia estandarizada entre grupos sobre estos parámetros, (e.g., entre el τ del ítem k en el grupo 1 y el τ del ítem k en el grupo 2), cuyo resultado se valora como la *d* (Cohen, 1988), y es compatible con los procedimientos mencionados anteriormente vinculados a medias latentes (Choi et al, 2009; Hancock, 2001).

Por último, los residuales, considerándolos como la *proporción de varianza no explicada por el constructo*, son comparados usando el estimador *h* (Cohen, 1988) siguiendo puntos de corte establecidos (.20, diferencia pequeña; .50, moderada; .80, grande), orientado a la comparación de proporciones. Para una revisión técnica de las expresiones matemáticas involucradas se recomienda la lectura de Pornprasertmanit (2014).

Aunque usualmente se toma el criterio racional de Cohen (1988) sobre los puntos de corte para decir cualitativamente (trivial, pequeño, mediano, grande), la magnitud de la diferencia en estos parámetros los debe elegir el usuario con base empírica o racional, evaluando la generalidad de los mismos a su situación de investigación (Merino-Soto & Copez-Lonzoy, 2018).

Presentación del programa

Existe una *función* elaborada para el entorno R que se ejecuta con el paquete *semTools* (Pornprasertmanit, 2014). El programa es de libre acceso, y se aplica a ítems continuos o categóricos para obtener el ES en cada parámetro (e.g., cargas factoriales, interceptos o *thresholds*, y residuales). Sin embargo, su mayor desventaja es que requiere conocimientos de R, cuya curva de aprendizaje es generalmente lenta para el usuario habituado a programas estadísticos basados en una interfase visual o cuadros de diálogo.

Para tener un medio más amigable para el usuario, se ha creado el módulo *ESinvariance* en MS Excel®, disponible sin costo al lector interesado, elaborado con base en la información derivada de Pornprasertmanit (2014). El módulo posee una sola ventana, en la cual se solicita al usuario la información del GR y GF. Los datos necesarios para la comparación son: tamaño muestral (n), media y desviación estándar (ítem), cargas factoriales, e interceptos (o *thresholds*, según corresponda), los que pueden derivarse del software que se utiliza más frecuentemente (e.g., AMOS, EQS, Mplus, LISREL, entre otros) y se deben colocar directamente en el módulo.

El tamaño muestral (n) debe colocarse una sola vez, ya que es común a todos los ítems. Sin embargo, la media y DE del ítem, así como las cargas factoriales (λ), interceptos (ν), y *thresholds* (τ) deben colocarse por cada ítem de forma independiente. Los residuales (θ) se calculan automáticamente ($\theta = 1-\lambda^2$). Repetir el proceso para el GF.

Una vez que son introducidos los datos al módulo, aparecen automáticamente las ES para la comparación de cargas factoriales, interceptos (o *thresholds*, según corresponda), y residuales. Para el caso de los *thresholds*, se debe seleccionar aquellos resultados compatibles con el número de categorías de respuesta de los ítems (por ejemplo, si tiene tres opciones de respuesta, se consideran dos *thresholds*). Una ES con signo negativo indica que el parámetro (e.g., carga factorial) es mayor en el GF.

El módulo puede ser solicitado de forma gratuita al primer autor del manuscrito.

Aplicación

Para ejemplificar el procedimiento se utilizaron los resultados y la base de datos de un estudio previo sobre una medida de agotamiento emocional (Dominguez-Lara, Fernández-Arata, Manrique-Millones, Alarcón-Parco,

& Díaz-Peñaloza, 2018). El agotamiento emocional académico es la sensación de cansancio, fatiga y falta de energía que experimenta el estudiante con relación a sus estudios, y se constituye como primera etapa y núcleo del burnout académico (Caballero, Hederich, & Palacio, 2010). En este estudio, con base el método de máxima verosimilitud robusta (ML-R) y matrices de covarianzas, se concluyó que una medida de agotamiento emocional académico tenía un grado suficiente de invarianza entre varones (GR; n=253) y mujeres (GF; n=849). Para utilizar el módulo se obtuvieron datos correspondientes a la *invarianza configural*: medidas descriptivas (e.g., media del ítem), así como las cargas factoriales, interceptos y residuales (Tabla 1).

Tabla 1
Cargas, interceptos, thresholds y residuales de los ítems en los grupos de referencia y focal

	Grupo de referencia (varones)								
	Cargas factoriales		Interceptos	Residuales		Thresholds			
	λ_{MLR}	λ_{WLSMV}	ν	θ_{MLR}	θ_{WLSMV}	τ_1	τ_2	τ_3	τ_4
Ítem 1	.626	.664	.660	.608	.559	-1.081	.020	1.276	2.414
Ítem 2	.451	.495	1.056	.797	.755	-.823	.143	1.154	1.878
Ítem 3	.616	.676	.317	.621	.543	-.488	.445	1.660	2.521
Ítem 4	.515	.555	1.114	.735	.693	-1.110	-.197	.799	2.057
Ítem 5	.648	.711	.183	.580	.495	-.407	.392	1.652	2.817
Ítem 6	.739	.786	.424	.454	.382	-1.422	-.331	1.264	2.703
Ítem 7	.705	.752	.199	.503	.434	-.900	.419	1.547	2.819
Ítem 8	.655	.701	.562	.571	.509	-1.232	-.230	.864	1.839
Ítem 9	.734	.771	.150	.461	.405	-1.144	.008	1.045	2.351
Ítem 10	.771	.812	.261	.406	.340	-1.691	-.333	1.078	2.133
	Grupo focal (mujeres)								
	Cargas factoriales		Interceptos	Residuales		Thresholds			
	λ_{MLR}	λ_{WLSMV}	ν	θ_{MLR}	θ_{WLSMV}	τ_1	τ_2	τ_3	τ_4
Ítem 1	.586	.640	.898	.657	.590	-1.440	-.457	.703	1.881
Ítem 2	.465	.498	.930	.784	.752	-.880	.039	1.049	1.864
Ítem 3	.575	.616	.220	.670	.621	-.410	.462	1.439	2.376
Ítem 4	.522	.553	1.047	.728	.694	-1.222	-.414	.620	1.705
Ítem 5	.655	.709	.006	.571	.497	-.728	.199	1.145	2.232
Ítem 6	.737	.779	.423	.457	.393	-1.791	-.530	.951	2.464
Ítem 7	.803	.844	-.217	.355	.288	-1.307	.179	1.797	3.338
Ítem 8	.713	.743	.417	.492	.448	-1.589	-.459	.911	2.171
Ítem 9	.740	.793	.173	.452	.371	-1.616	-.459	.755	2.114
Ítem 10	.738	.766	.113	.455	.413	-1.458	-.321	.772	2.079

Nota: λ_{MLR}=cargas factoriales con el método ML-Robusto; λ_{WLSMV}=cargas factoriales con el método WLSMV; ν=intercepto; θ_{MLR}=residuales con el método ML-Robusto; θ_{WLSMV}=residuales con el método WLSMV; τ=*threshold*

Para propósitos de usar el módulo *ESinvariance* con la presunción que los ítems son variables categóricas, complementariamente se realizó un análisis aplicando el estimador *mínimos cuadrados ponderados con varianza ajustada* (WLSMV; por sus siglas en inglés) y matrices policóricas para obtener los *thresholds*, además de las cargas factoriales y residuales (Tabla 1).

Las magnitudes de las ES brindan indicadores compatibles con las conclusiones iniciales del estudio de Dominguez-Lara et al. (2018), es decir, que la medida es invariante según el género (Tabla 2); sin embargo, algunos ítems presentaron resultados interesantes. Por ejemplo, en cuanto a las cargas factoriales, la diferencia entre los grupos fue trivial (q<.10) independientemente del método de estimación empleado (ML-Robusto o WLSMV). Esto quiere decir que el constructo es representado de forma similar entre los varones y mujeres.

Por otro lado, si bien los interceptos de los ítems 1 y 7 presentan diferencias pequeñas, a favor de las mujeres y varones respectivamente, que sugieren falta de invarianza (método ML-Robusto). Esto podría interpretarse, para el caso del ítem 7 (*Me siento emocionalmente agotado por mis estudios*), de la siguiente manera: un varón con determinado *nivel verdadero* en agotamiento emocional obtendría

una puntuación más alta que una mujer con el mismo *nivel verdadero* en dicho rasgo. Sin embargo, a nivel de *thresholds* (método WLSMV) las magnitudes de *d* fueron triviales (<.20), llegando incluso a ser insignificantes (e.g., ítem 7).

Finalmente, en cuanto a los residuales, solo la ES asociada al ítem 7 es elevada con ambos métodos (>.20), lo que indicaría que el error de medición es distinto en ambos grupos, es decir, que fuentes ajenas al constructo afectan de forma diferencial las respuestas de varones y mujeres. Con todo, podría considerarse una medida invariante aún con esos hallazgos debido a que la cantidad de parámetros no-invariantes no fue significativa (<20%; Dimitrov, 2010).

Tabla 2
Medidas de magnitud del efecto (ES)

	ES para cargas factoriales (λ)		ES para interceptos (ν)	ES para los thresholds (τ)				ES para los residuales (θ)	
	ES-λ_{MLR} (*q*)	ES-λ_{WLSMV} (*q*)	(*d*)	ES- τ_1 (*d*)	ES- τ_2 (*d*)	ES- τ_3 (*d*)	ES- τ_4 (*d*)	ES- θ_{MLR} (*h*)	ES- θ_{WLSMV} (*h*)
Ítem 1	.029	.016	**-.214**	.107	.142	-.171	-.159	.101	.063
Ítem 2	-.011	-.002	.112	.021	.038	-.038	-.005	-.032	-.007
Ítem 3	.029	.041	.087	-.027	-.006	-.078	-.051	.102	.157
Ítem 4	-.005	.002	.058	.038	.073	-.060	-.118	-.016	.005
Ítem 5	-.005	.001	.147	.106	.064	-.168	-.194	-.019	.006
Ítem 6	.001	.004	.001	.087	.047	-.074	-.057	.006	.023
Ítem 7	-.060	-.052	**.379**	.092	.054	.056	.117	**-.300**	**-.307**
Ítem 8	-.038	-.026	.126	.099	.064	.013	.092	-.160	-.122
Ítem 9	-.004	-.013	-.019	.128	.127	-.079	-.064	-.018	-.071
Ítem 10	.020	.026	.122	-.065	-.003	-.085	-.015	.101	.150

Nota: λ_{ML}R=cargas factoriales con el método ML-Robusto; λ_{WLSMV}=cargas factoriales con el método WLSMV; ν=intercepto; θ_{MLR}= residuales con el método ML-Robusto; θ_{WLSMV}=residuales con el método WLSMV; τ=*threshold*. En negrita=magnitud de la ES que sugiere falta de invarianza

Conclusiones

El análisis de la IM es un tópico recurrente en la investigación instrumental e implementar medidas de ES es necesario debido a las limitaciones que acarrea el enfoque basado solo en los *índices de modificación*. Entonces, analizar los parámetros individualmente aportaría información para una mejor comprensión de la evaluación de determinado constructo en diversos grupos, tanto en lo concerniente al grado de representatividad de los ítems (e.g., cargas factoriales), punto de partida respecto al rasgo evaluado (e.g., interceptos/*thresholds*) o cantidad de error asociado a la medición (e.g., residuales).

Referencias

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*, 3-25. doi: 10.1037/amp0000191.

Caballero, C, Hederich, C, & Palacio, J. (2010). El burnout académico: delimitación del síndrome y factores asociados a su aparición. *Revista Latinoamericana de Psicología, 42*(1), 131-146.

Cheung, G.W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255. Doi: 10.1207/S15328007SEM0902_5

Choi, J., Fan, W., & Hancock, G. R. (2009). A note on confidence intervals for two-group latent mean effect size measures. *Multivariate Behavioral Research, 44*(3), 396-406. doi: 10.1080/00273170902938902

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

DeLong, D. & Elbeck, M. (2018). An Exploratory Study of the Influence of Soft and Hard Skills on Entry Level Marketing Position Interviews. *Marketing Education Review, 28*(3), 159-169. doi: 10.1080/10528008.2017.1349475

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149. doi: 10.1177/0748175610373459

Dominguez-Lara, S., Fernández-Arata, M., Manrique-Millones, D., Alarcón-Parco, D., Díaz-Peñaloza, M. (2018). Datos normativos de una escala de agotamiento emocional académico en estudiantes universitarios de psicología de Lima (Perú). *Educación Médica, 19*(3), 246 - 255. doi: 10.1016/j.edumed.2017.09.002

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha coefficient is the same for two tests administered to the same sample. *Psychometrika, 49*, 99-105.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*(3), 373-388. doi: 10.1007/BF02294440

Merino-Soto, C. & Copez-Lonzoy, A. (2018). Un problema con diferentes rostros: arbitrarios niveles de tamaño del efecto. *Enfermería Clínica, 28*(6), 347 - 404. doi: 10.1016/j.enfcli.2018.05.001

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology, 60*, 65-82. doi: 10.1016/j.jsp.2016.11.002

Penfield, R. D., & Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213-225.

Pornprasertmanit, S. (2014). *A Note on Effect Size for Measurement Invariance.* Recuperado desde: http://cran.irsn.fr/web/packages/semTools/vignettes/partialInvariance.pdf

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371-384. doi: 10.1007/BF02294623

recebido em junho de 2017
aprovado em setembro de 2019

Sobre os autores

Sergio Dominguez-Lara es Psicólogo y Doctor en Psicología. Docente del Instituto de Investigación de Psicología de la Universidad de San Martín de Porres (Lima, Perú).

César Merino-Soto es Psicólogo y Magíster en Psicología. Doctorando en la Universidad Autónoma del Estado de Morelos (México). Docente del Instituto de Investigación de Psicología de la Universidad de San Martín de Porres (Lima, Perú).