XII CONGRESO DE METODOLOGÍA DE LAS CIENCIAS SOCIALES Y DE LA SALUD

DONOSTIA 19-22 DE JULIO DE 2011

LIBRODEACTAS

Universidad del País Vasco
Euskal Herriko Unibertsitatea

XII CONGRESO DE METODOLOGÍA DE LAS CIENCIAS SOCIALES Y DE LA SALUD

XII CONGRESO DE METODOLOGÍA DE LAS CIENCIAS SOCIALES Y DE LA SALUD

Comité de Honor

Rector Magnífico de la UPV/EHU
Don Iñaki Goirizelaia

Sra. Decana de la Facultad de Psicología de la UPV/EHU
Dña. Ana Isabel Vergara

Sr. Director del Dpto. de Psicología Social y Metodología de las Ciencias del Comportamiento de la UPV/EHU
Don Juan José Arrospide

Sr. Presidente de la European Association of Methodology
Don José Muñiz

Sra. Presidenta de la Asociación Española de Metodología de las Ciencias del Comportamiento
Dña. Teresa Anguera

Comité Científico

Presidente: Dr. Jaume Arnau (Universidad de Barcelona)
Dr. Julio Olea (Universidad Autónoma de Madrid)
Dra. Dolors Riba (Universidad Autónoma de Barcelona)
Dra. Gloria Seoane (Universidad de Santiago de Compostela)

Comité Organizador

Presidencia: Dra. Nekane Balluerka y Dra. Esther Torres
Tesorería: Dra. Ana Isabel Vergara
Vocales: Dra. Arantxa Gorostiaga
Dra. Izaskun Ibabe
Dr. Xabier Isasi
Dña. Nerea Lertxundi
Dra. Laura Vozmediano

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

ARGITALPEN ZERBITZUA
SERVICIO EDITORIAL

Edita: Asociación Española de Metodología de las Ciencias del Comportamiento

Coordinadoras: Nekane Balluerka, Arantxa Gorostiaga, Nerea Lertxundi y Laura Vozmediano

ISBN: 978-84-9860-633-1
Depósito Legal: BI 432-2012

Fotocomposición / Fotokonposizioa: Rali, S.A.
Particular de Costa, 12-14 - 48010 Bilbao

ÍNDICE

SIMPOSIOS

SESIONES PARALELAS

APLICACIONES PSICOMÉTRICAS

CUESTIONES TEÓRICAS

MÉTODOS ESTADÍSTICOS

EXPERIENCIAS DOCENTES

APLICACIONES METODOLÓGICAS

SIMPOSIOS

STATISTICAL MODELING WITH ROBUST PROCEDURES. STUDYING THE AWKWARD ASSUMPTIONS AND OTHER STRATEGIES OF DATA COLLECTION

Coordinadores: Maribel Peró y Joan Guàrdia
Universidad de Barcelona

In this symposium we try to establish some contribution to the statistics modelization when the assumptions are violated. Usually, there a lot of violations when a multivariate statistical technique is used, but maybe, it is more usual that the researchers don't verify these assumptions and the journals don't ask about the achievement of the assumptions of multivariate statistical modelization. In these symposium we don't try to analyze the correct use of multivariate statistical techniques in relation to their assumptions, maybe this is related to the irresponsible use of statistical software and the lost of statistical training in the researchers that generally search a quick solution to their analysis. When the statistical techniques are more complex is more habitual that the researchers don't achieve the assumptions and use the statistical software in a irresponsible way. There are few works, where the authors present information in relation to the achievement of statistical assumptions. Homocedasticity, esphericity test, normality or multinormality, initial matrix determinant, lineal dependency between exogenous variables, observed residuals distribution, symmetry, sample size, etc. aren't object of researchers attention in their analysis. In this symposium we try to show some contribution that could solve some of these problems. First of all, we deal with the effect of some aspects (symmetry, sample size) in the classical and bayessian estimation in logistic regression; the second one could offer a better solution that the first one when we work with small samples and biased distributions. After, we present empirical evidence of the use of robust estimators in the two independent groups comparison, basically using the median. We will follow with the contribution to the use of textual analysis techniques, showing empirical evidences of their use and recommendations use. Finally, we would try to give some new algorithms derived of Exploratory Structural Equation Modeling, generated in order to solve the violation that the researchers make in the use of confirmatory models as an exploratory technique.

KEY WORDS: Symulation, Statistics modelization, Multivariate statistics, Qualitative analysis.

ESTIMACIÓN CLÁSICA Y BAYESIANA EN MODELOS DE REGRESIÓN LOGÍSTICA. UN ESTUDIO DE SIMULACIÓN

Amalia Gordóvil, Joan Guàrdia y Maribel Peró
Universidad de Barcelona
Correo electrónico: amalia.gordovil@ub.edu

Resumen

El presente trabajo compara estimaciones clásicas y bayesianas en modelos de regresión logística en casos de tamaños muestrales inferiores a los deseables. Para abordar dicho objetivo se han programado una serie de simulaciones en las que se varían condiciones como el tipo de variable, valores distribucionales y valores de asimetría. En el caso de las estimaciones bayesianas, se incorporó por defecto un *prior* del tipo «algo informativo» (weakly informative). El método de estimación utilizado para estimaciones clásicas y bayesianas fue el mismo: mínimos cuadrados ponderados iterativos. Futuras investigaciones donde se trabaje con diferentes tipos de *priors* podrían contribuir a un estudio más exhaustivo del problema.

La inexorable relación entre tamaño de muestra y precisión en la estimación, resulta ampliamente conocida en el ámbito de la investigación. Por ello, tipos de inferencias que son adecuados en trabajos con tamaños muestrales grandes resultan inapropiados cuando aplicamos modelos de regresión logística (LR) a muestras pequeñas (Potter, 2005).

En este trabajo tenemos por objetivo comparar estimaciones clásicas y bayesianas en modelos de RL con muestra pequeña. Para ello, hemos definido rutinas de simulación donde establecemos diferentes tipos de variables, distintos valores distribucionales y variamos condiciones de asimetría.

MÉTODO

Se definieron diferentes condiciones para la variable dependiente (VD) y para dos variables independientes (VI). Estas tres variables conformaron modelos de RL. Primero se manipularon ciertos parámetros de las rutinas de simulación para después analizar el efecto sobre los modelos de RL. Dichos resultados se analizaron en términos de predicciones correctas sobre la VD y valores de coeficientes y errores estándar de las VIs. En este estudio, se presentan resultados correspondientes a predicciones correctas sobre la VD.

Se establecieron dos condiciones para la VD definidas por dos valores en la distribución binomial: π_1=.5 (primera condición) y π_1=.2 (segunda condición). Es decir, se simuló una población con presencia de trastorno en un 50% de los casos y ausencia en el 50% restante (primera condición para VD: π_1=.5; π_0=.5). Asimismo, se simuló una segunda población con presencia de trastorno en el 20% de los casos y ausencia en el 80% (segunda condición para VD: π_1=2; π_0=.8). El subíndice 1 hace referencia al grupo con presencia de trastorno mientras que el subíndice 0 denota al grupo con ausencia de trastorno.

En referencia a las dos VIs, una fue definida como binaria y otra como continua. Respecto a la VI binaria, se establecieron dos condiciones a partir de distribución binomial: 1) condición de no relación (π_1=.5; π_0=.5; recuerde que los subíndices se refieren a los grupo definidos por la VD) y 2) condición de relación (π_1=.7; π_0=.4). En la condición de no relación se simuló la presencia de un factor de exposición en el 50% de los casos y ausente en el restante 50% tanto en el grupo de presencia como en el de ausencia de trastorno. Por tanto, el factor de exposición no estaba relacionado con la presencia del trastorno. En cuanto a la condición de relación, el factor de exposición estuvo presente en el 70% de los casos y ausente en el restante 30% de casos en el grupo con presencia de trastorno. Respecto al grupo de ausencia de trastorno, el factor de exposición se definió como presente en el 40% de casos y ausente en el restante 60%. Por tanto, dicho factor de exposición estuvo relacionado con la presencia de trastorno.

Por lo que se refiere a las VIs, se combinaron condiciones de relación y asimetría. En primer lugar, se definieron variables con distribución normal estandarizada (μ=0, σ^2=1). La asimetría fue generada utilizando la siguiente fórmula (Jiménez & Martínez, 2006),

$$Y = T_{g,h}(Z) = \frac{e^{g \cdot z} - 1}{g} \cdot e^{h \cdot z^2/2}$$

dónde *Z* es una variable con distribución normal, *g* es el valor de asimetría distribucional (0 indica simetría) y *h* indica la curtosis distribucional (se igualó a 0). Se fijó un valor de .8 para *g* generando así distribuciones asimétricas positivas:

$$Y = T_{g,0}(Z) = \mu + \sigma * \frac{e^{0.8 \cdot z} - 1}{0.8}$$

A su vez, se especificaron valores para media y desviación típica dado que la variación en los valores de la media fue utilizada para definir condiciones de relación y no relación. Las condiciones de no relación fueron definidas como aquéllas en las que las medias del grupo con presencia de trastorno (μ_1) presentó valores similares a las medias del grupo con ausencia de trastorno (μ_0). El valor de la media se estableció en 100. Por tanto, dado que el valor de la media fue igual para los dos grupos, éste valor permaneció no relacionado con la presencia o ausencia de trastorno. Por otro lado, se definieron como condiciones de relación aquéllas en las que

las medias del grupo de presencia de trastorno (μ_1) difirieron con las correspondientes al grupo donde el trastorno permaneció ausente (μ_0). Los valores de las medias se establecieron así: μ_1=100, μ_0=120. Por tanto, si los valores de medias son distintos en ambos grupos, podemos decir que dichos valores están relacionados con la presencia del trastorno. Teniendo todo esto en cuenta, las condiciones de no relación fueron: 1) $\mu_1=\mu_0$=100; σ=15; g_1=.0 y 2) $\mu_1=\mu_0$=100; σ=15; g_1=.8. Las CR fueron: 3) μ_1=100; μ_0=120; σ=15; g_1=.0 y 4) μ_1=100; μ_0=120; σ=15; g_1=.8. Tal y como puede observarse, la asimetría positiva se generó en el grupo 1 y los valores de desviación estándar permanecieron constantes.

Cabe puntualizar que las condiciones de no relación y las condiciones de relación se refieren a la relación establecida entre VD (presencia o ausencia de trastorno) y VIs (presencia de factor de exposición).

En cada condición simulada se estableció un tamaño muestral de 100 y se realizaron 10000 iteraciones. El conjunto de simulaciones programadas aparece esquematizado en la tabla 1.

Tabla 1. Esquema de Condiciones de Simulación

Tipo de VI	**Tipo de Condición**	**Valores Distribucionales**	
		VD	**VI**
Binaria	No relación	$\pi1$=.5;$\pi0$=.5	$\pi1$=.5;$\pi0$=.5
		$\pi1$=.2;$\pi0$=.8	
	Relación	$\pi1$=.5;$\pi0$=.5	$\pi1$=.7;$\pi0$=.4
		$\pi1$=.2;$\pi0$=.8	
Continua	No relación	$\pi1$=.5;$\pi0$=.5	m1=m0=100;σ=15;g1=.0 m1=m0=100;σ=15;g1=.8
		$\pi1$=.2;$\pi0$=.8	
	Relación	$\pi1$=.5;$\pi0$=.5	m1=100;m0=120;σ=15;g1=.0
		$\pi1$=.2;$\pi0$=.8	m1=100;m0=120;σ=15;g1=.8

El método de estimación utilizado para ambas estimaciones (clásicas y bayesianas) fue el de mínimos cuadrados ponderados iterativos. Se trata de una simplificación del algoritmo de máxima verosimilitud correspondiente a la familia distribucional de funciones exponenciales. Una ventaja derivada de utilizar este método frente al de máxima verosimilitud consiste en una mayor facilidad de procesamiento del modelo. El lector interesado puede remitirse a Hilbe (2009) para una explicación detallada de métodos de estimación en modelos de RL.

En las estimaciones bayesianas se introdujo un *prior* del tipo «algo informativo» con el objetivo de valorar si éste podía contribuir a la detección de diferencias entre

estimaciones clásicas y bayesianas. Siguiendo a Gelman, Jakulin, Pittau y Su (2008), los modelos de RL fueron estimados realizando una simple adaptación del método de mínimos cuadrados ponderados iterativos. Se utilizó la distribución de Cauchy como *prior* distribucional. Ésta distribución a priori, se construyó mediante el escalamiento de las variables no binarias con valores de media 0 y deviación típica .5 y el escalamiento de variables binarias con valores de media 0 y diferencias de 1 en valores mínimo y máximo. Tras obtener variables estandarizadas, la distribución a priori independiente de Cauchy se asignó a los coeficientes en modelos de RL (a excepción de la constante del modelo) (ver Gelman et al., 2008 para una explicación exhaustiva de la construcción del *prior*). De acuerdo con los autores, la inclusión de este tipo de información puede contribuir a la regularización de inferencias extremas que se obtendrían en caso de utilizar *priors* no informativos. Vea que, cuando nos referimos a estimaciones bayesianas, hacemos alusión a la introducción del *prior* algo informativo basado en la distribución de Cauchy. Cuando nos referimos a estimaciones clásicas, ningún *prior* fue introducido en el modelo.

Se obtuvieron estimaciones clásicas y bayesianas para cada condición simulada. Para las estimaciones clásicas se utilizó la función *glm* del paquete *stats* (R Development Core Team, 2010) mientras que para las estimaciones clásicas se recurrió a la función *bayesglm* del paquete *arm* (Gelman et al., 2010). Se realizaron pruebas de comparación de proporciones (muestras relacionadas) y test binomial para identificar diferencias significativas entre estimaciones clásicas y bayesianas respecto al porcentaje de p-valores correctos. Para todo ello se utilizó el entorno R 2.11.1 (R Development Core Team, 2010).

RESULTADOS

El análisis de diferencias entre estimaciones clásicas y bayesianas sobre porcentajes de decisiones correctas se realizó mediante pruebas de comparación de proporciones para muestras relacionadas (test binomial en casos de incumplimiento de condiciones de aplicación) obteniéndose tests unilaterales suponiendo un mejor resultado para las estimaciones bayesianas. En la tabla 2 se muestran resultados para ambas estimaciones (clásica y bayesiana). Los resultados se presentan ordenados según las condiciones de simulación. Es decir, la primera y segunda condición de la VI binaria corresponden a condiciones de no relación; la tercera y cuarta condición de la VI binaria son condiciones de relación. En cuanto a la VI continua, las cuatro primeras condiciones mostradas en la tabla 2 son condiciones de no relación, mientras que las cuatro últimas son condiciones de relación.

Respecto a la VI binaria, ambas estimaciones detectan correctamente las condiciones de no relación (porcentajes de p-valores correctos entre 96.25 y 97.35) existiendo diferencias significativas a favor de las estimaciones bayesianas ($z=14.34$, IC=.08-.92, $p<.001$ para la distribución de .5 en la VD y VI; $z=16.98$, IC=.09-.91, $p<.001$ para la distribución .2 en la VD y .5 en la VI).

Tabla 2. Porcentaje de p-valores Correctos para Estimaciones Clásicas y Bayesianas

Tipo de VI	Valores Distribucionales	Tipo de Estimación		Test Comparación Proporciones			Test Binomial
		Clásica	Bayesiana	z	IC	p	p
Binaria	VD=.5;VI=.5	96.25	96.85	14.34	.08-.92	<.001	
	VD=.2;VI=.5	96.48	97.35	16.98	.09-.91	<.001	
	VD=.5,VI=.7/.4	54.23	53.23	13.39	.27-.73	<.001	
	VD=.2,VI=.7/.4	41.83	39.70	21.45	.23-.77	<.001	
Continua	VD=.5;m1=m0=100;S	95.25	95.95				<.001
	VD=.5;m1=m0=100;A	35.90	0				<.001
	VD=.2;m1=m0=100;S	95.70	96.45				<.001
	VD=.2;m1=m0=100;A	34.65	0				<.001
	VD=.5;m1=100;m0=120;S	100	100				<.001
	VD=.5;m1=100;m0=120;A	100	100				1
	VD=.2;m1=100;m0=120;S	99.85	99.75				1
	VD=.2;m1=100;m0=120;A	100	100				1

Nota: S, simetría; A, asimetría.

En el caso de condiciones de relación de la VI binaria los porcentajes de p-valores correctos disminuyen considerablemente (entre 39.70% y 54.23%), siendo mejores las estimaciones clásicas (aunque no superan porcentajes correctos mayores al 54.30%). En cuanto a porcentajes de p-valores correctos para la VI continua en condiciones de relación, se obtienen resultados correctos y similares en ambas estimaciones. Tal y como puede apreciarse en la tabla 2, se logra un porcentaje del 100% de p-valores correctos en todas las situaciones excepto el 99.85% conseguido por estimaciones clásicas y el 99.75% logrado por estimaciones bayesianas en el caso de valores distribucionales de .2 en la VD y VI asimétrica con valores de media 100 en el grupo de presencia de trastorno y 120 en el grupo de ausencia de trastorno. Respecto a condiciones de no relación, las estimaciones bayesianas superan a las estimaciones clásicas en casos de VI simétricas con distribución de .5 y .2 de la VD ($p<.001$ en ambos casos). Sin embargo, existen importantes problemas para ambas estimaciones bajo condiciones de no relación cuando se genera asimetría a la VI continua.

DISCUSIÓN

El presente trabajo compara resultados entre dos tipos de estimaciones en modelos simulados de RL. Una característica definitoria de los métodos bayesianos es que no únicamente se basan en informaciones referentes a la muestra actual, sino que permiten la incorporación de informaciones existentes a priori, en relación con el objeto de estudio. En este trabajo, denominamos estimaciones bayesianas al hecho de introducir un *prior* algo informativo en los modelos simulados de RL. Se trata de una distribución que se utilizará por defecto, propuesta por Gelman et al (2008). En las estimaciones clásicas presentadas aquí, no se ha introducido ningún tipo de *prior* en los modelos a estimar.

Se han generado modelos de RL compuestos por dos VI*s* (una binaria y otra continua). En los procesos de simulación, se establecieron modelos en los que los valores permanecían relacionados (condiciones de relación) y modelos en los que los valores no estaban relacionados (condiciones de no relación).

Respecto a la VI binaria, el porcentaje de p-valores correctamente detectados fue similar en ambas estimaciones bajo condiciones de no relación. Sin embargo, el porcentaje fue menor bajo condiciones de relación. Respecto a la VI continua, ambas estimaciones detectaron porcentajes similares de p-valores correctos bajo condiciones de relación. En condiciones de no relación los resultados fueron menos satisfactorios, algo particularmente problemático en casos de asimetría positiva.

Una limitación importante del estudio es el trabajo con un único tamaño muestral, algo a ampliar en futuros estudios. Otra posible mejoría sería la incorporación de diferentes valores distribucionales del factor de exposición simulado. Aquí se ha trabajado con dos condiciones: una distribución idéntica (π_1=.5; π_0=.5) y una distribución diferente (π_1=.7; π_0=.4) del factor de exposición en ambos grupos (grupo de presencia y grupo de ausencia de trastorno). Extremar las diferencias entre las distribuciones del factor de exposición en los grupos de trastorno y no trastorno podría contribuir a un análisis más exhaustivo. Otro aspecto a tener en cuenta es la inclusión de diferentes tipos de *priors* (e.g., no informativos, muy informativos). El *prior* con el que hemos trabajado (algo informativo) puede utilizarse en un amplio rango de situaciones (Gelamn et al., 2008). Sin embargo, este *prior* utilizado por defecto, podría constituir una línea base sobre la cual podrían utilizarse *priors* informativos.

En resumen, una primera línea de trabajo consiste en estudiar resultados procedentes de modelos de RL con muestra pequeña basados en datos reales en los cuales se comparen diferentes tipos de *priors* (e.g. priors informativos generados a partir de *background* relevante sobre la materia). Otro aspecto sería la simulación de modelos de RL con muestra pequeña trabajando con *priors* no informativos, algo informativos y muy informativos en comparación a estimaciones clásicas. Todo ello, sin olvidar la necesidad de mejorar los problemas existentes en torno a distribuciones asimétricas, no poco frecuentes en el ámbito de la psicología.

NOTA DE LOS AUTORES

Estudio subvencionado por el Comissionat per a Universitats I Recerca, Generalitat de Catalunya (Departament de Innovació, Universitats I Empresa), Fons Social Europeu.

REFERENCIAS

Gelman, A., Jakulin, A., Pittau, M.G., & Su, Y.S. (2008). A weakly default prior distribution for logistic and other regression models. *The Annals of Applied Statistics,* 2, 1360-1383.

Gelman, A., Su, Y.S., Yajima, M., Hill, J., Pittau, M.J., Kerman, J., & Theng, T. (2010). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models (Version 1.3-08) [Computer software]. Retrieved from http://CRAN.R-project.org/package=arm.

Hilbe, J. M. (2009). *Logistic regression models*. Florida: Taylor & Francis.

Jiménez, J.A., & Martínez, G. (2006). Una estimación del parámetro de la distribución g de Tukey. *Revista Colombiana de Estadística*, 29, 1-16.

Potter, D.M. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Statistics in Medicine,* 24, 693-708.

R Development Core Team. (2010). R: A Languaje and Enviroment for Statistical Computing (Version 2.11.1) {Computer software}. Viena: R Foundation for Statistical Computing.

ESTUDIO DE LA EFECTIVIDAD DE DIFERENTES MÉTODOS ROBUSTOS PARA LA DECISIÓN ESTADÍSTICA EN LA COMPARACIÓN DE DOS GRUPOS INDEPENDIENTES

Maribel Peró-Cebollero y Joan Guàrdia-Olmos
Universidad de Barcelona
Correo electrónico: mpero@ub.edu, jguardia@ub.edu

Resumen

En el presente trabajo se analiza la bondad de técnicas robustas en la decisión estadística en la comparación de dos grupos independientes, básicamente a partir de la comparación de intervalos de confianza de medianas. Se ha realizado un estudio de simulación utilizando el software R y manipulando la existencia o no de diferencias entre los dos grupos, el tamaño de muestra y el grado de simetría. Los resultados muestran una adecuada sensibilidad de los intervalos de confianza basados en la mediana en muestras grandes (superiores a 30) y una adecuada especificidad. De todos modos el estadístico con menor tasa de error es la t de Yuen-Welch.

Es común en el ámbito de la investigación aplicada en psicología, y en general en las ciencias sociales, el uso de la estadística más clásica en las pruebas de decisión. Aunque cada vez existen más controles al respecto sobre la adecuación de la estadística paramétrica en las investigaciones aplicadas, y con mayor frecuencia se pueden ver publicaciones en que se utiliza estadística no paramétrica ante el incumplimiento de los supuestos y condiciones básicas de las diferentes pruebas paramétricas, aún existe una enorme cantidad de trabajos en los que no se hace mención a estos supuestos o condiciones. En 1977 Tukey revolucionó el mundo de la estadística con la concepción del análisis exploratorio de datos (EDA), de todos modos, las ideas que aportaba Tukey en 1977 se han incorporado en la estadística descriptiva, pero este paso no se ha llevado a cabo en la estadística infencial. Por otra parte, cabe comentar, que con mayor asiduidad en los últimos tiempos se está recomendando que en las publicaciones científicas, el grado de significación asociado al estadístico de contraste aparezca acompañado de indicadores de tamaño del efecto de la relación y intervalos de confianza (APA, 2001; Cumming y Finch, 2001; o Wilkinson y the Task Force for Statistical Inference, 1995). Pero pocos son los trabajos que usan la comparación de los intervalos de confianza en la decisión estadística, como ejemplo se pueden citar los trabajos de Bonnet y Price (2002) o Cumming y Maillardet (2006). El bajo uso de los intervalos de confianza en la decisión estadística, posiblemente es debido a la falta de trabajos sobre la adecua-

ción de los mismos. Es por este motivo que en el presente estudio se pretende evaluar la bondad del uso de intervalos de confianza robustos, básicamente basados en la mediana, en la decisión estadística para la comparación de dos grupos independientes.

OBTENCIÓN DE LOS INTERVALOS DE CONFIANZA

La distribución muestral de medias sigue el modelo de la ley normal, por lo que la obtención del intervalo de confianza de la media se obtiene aplicando este modelo a partir de la siguiente fórmula:

$$\bar{x} \pm t_{(\alpha,\nu)} \cdot \frac{\sigma}{\sqrt{n}}$$

El cálculo del intervalo de confianza de la media recortada se obtiene igual que el intervalo de confianza anterior, pero en este caso trabajando con la media recortada y la varianza winsorizada a partir del mismo porcentaje de datos que se han eliminado en la media recortada, así como una corrección del tamaño del muestra en función del porcentaje de datos recortados (Wilcox, 2005). Por tanto la fórmula es:

$$\bar{x}_t \pm t_{(\acute{a}\ 2,\nu)} \cdot \frac{s_w}{(1-2\gamma)\sqrt{n}}$$

En el caso de la mediana se han utilizado cinco métodos para la obtención del intervalo de confianza, el basado en el error estándar, el basado en la distribución binomial, el basado en el error estándar a partir del método de McKean y Schraeder, el basado en el error estándar a partir del método de Marizt y Jarret y el basado en la estimación adaptativa kernel (Wilcox, 2005).

Para la obtención del intervalo de confianza de medianas basado en el error estándar (Kendall, 1945 o Mothes y Torrens-Ibern, 1970) si la población es normal con media μ y desviación típica σ y la muestra suficientemente grande, la ley de probabilidad de la mediana tiende a ser una ley normal con las siguientes características:

$$E[mediana] = \mu \qquad VAR[mediana] = \frac{\pi}{2}\frac{\sigma^2}{n} \qquad EE(mediana) = 1.253 \cdot \sigma / \sqrt{n}$$

en consecuencia el intervalo de confianza es:

$$Md \pm t_{(\alpha,\nu)} \cdot 1.253 \cdot \frac{\sigma}{\sqrt{n}}$$

El segundo método, cálculo del intervalo de confianza de la mediana a partir de la distribución binomial, se basa en la obtención de las posiciones del límite

inferior y el límite superior del intervalo a partir de la aplicación de la distribución binomial, dado que el número de observaciones por debajo del centil k sigue esta ley con parámetros n y k (Bland, 2003 y DeCoster y Burchill, 2000) y la mediana es el punto central de la distribución(k = 0,5), los parámetros que la definen son:

Posición del valor de la mediana: $\frac{n+1}{2}$

Error estándar: $\sqrt{n \cdot p \cdot (1-p)}$

por tanto el intervalo es:

$$\frac{n+1}{2} \pm t_{(\alpha,\nu)} \cdot \sqrt{n \cdot p \cdot (1-p)}$$

Una vez obtenidas las posiciones se redondean al entero más próximo y finalmente se obtienen los valores de la distribución observada que ocupan dichas posiciones.

La expresión de cálculo para la obtención del intervalo de confianza de la mediana a partir de la estimación del error estándar propuesta por McKean y Schraeder es (Wilcox, 2005):

$$ES = \left(\frac{x_{(n-k-1)} - x_k}{2\, z_{0.995}}\right), \qquad donde\ k = \frac{n-1}{2} - z_{0.995}\sqrt{\frac{n}{4}}$$

Finalmente, la expresión de cálculo para la obtención del intervalo de confianza de la mediana a partir de la estimación del error estándar propuesta por Maritz y Jarret es (Wilcox, 2005):

$$\lambda\, x_{k+1} + (1-\lambda)\, x_k \div \lambda\, x_{n-k} + (1-\lambda)\, x_{n-k+1}$$

Donde: $I = \frac{\gamma_k - 1 - \alpha}{\gamma_k - \gamma_{k+1}}, \lambda = \frac{(n-k)\, I}{k + (n-2k)\, I}$

MÉTODO

Con el fin de estudiar la adecuación de los intervalos de confianza mencionados anteriormente se ha llevado a cabo un estudio de simulación utilizando el software R (R Development Core Team, 2010). En concreto se han generado dos situaciones a estudiar, una de igualdad de medias poblacionales en los dos grupos a comparar ($\mu_x = \mu_y = 100$) y otra de diferencias de medias poblacionales ($\mu_x = 100$ y $\mu_y = 115$). En cada una de las situaciones se ha fijado la desviación típica a 10 en las dos poblaciones a comparar y se ha trabajado con tres valores de tamaño de muestra (10, 30 y 50 pero iguales en los dos grupos comparados) y dos valores de asimetría aplicados a las dos poblaciones por tanto 4 condiciones diferentes ($g_x = 0$, $g_x = .8$, $g_y = 0$ y $g_y = .8$). Así pues para cada situación se han estudiado 12 condiciones distintas. Para cada condición se han generado 5000 muestras de cada

población a comparar. En la figura 1 se muestra gráficamente el esquema de la simulación realizada.

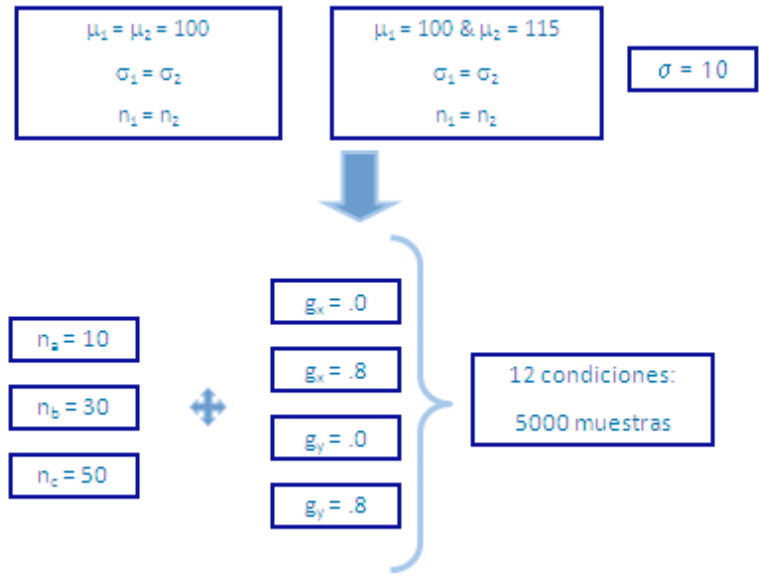


Figura 1. Representación gráfica del proceso de las simulaciones generadas

Cada una de las muestras se han generado a partir de una distribución normal estandarizada ($\mu = 0$ y $\sigma = 1$), posteriormente se ha utilizado la distribución *gh* con el fin de generar el nivel de asimetría deseado aplicando la fórmula siguiente:

$$\frac{e^{gh} - 1}{g} \cdot e^{h \cdot z^2/2}$$

Donde g indica la asimetría que se puede generar y h la curtosis que se puede generar en la distribución normal. El 0 en ambos casos indicaría una distribución perfectamente simétrica y mesocúrtica, en tanto que a medida que los valores de estos parámetros se alejan del 0 aumenta la asimetría y el grado de apuntamiento (Field y Genton, 2006 o Wilcox, 2005). Posteriormente se ha multiplicado los valores en cada muestra por la desviación típica de 10 y se ha sumado el valor de la media del grupo 100 o 115 dependiendo de la situación estudiada.

ANÁLISIS DE DATOS

Para cada par de muestras a comparar se ha obtenido el estadístico t de *Student* de grupos independientes junto con su condición de aplicación, la prueba no paramétrica U de Mann-Whitney, la t de Yuen-Welch (Wilcox, 2005), cuya fórmula de cálculo se muestra a continuación:

$$t = \frac{|\bar{x}_{t1} - \bar{x}_{t2}|}{\sqrt{d_1 + d_2}} \quad d_j = \frac{(n_j - 1)\, s_{wj}^2}{h_j\,(h_j - 1)} \quad h_j = n_j - 2\gamma\, n_j \quad \gamma - \textit{proporción de recorte}$$

También se han obtenido los diferentes intervalos de confianza mencionados anteriormente. El criterio de decisión en la comparación de los intervalos de confianza para determinar la existencia de diferencias entre los dos grupos ha consistido en la no inclusión del estadístico resumen en el grupo «x» en el intervalo de confianza generado en el grupo «y» y viceversa, se debían cumplir ambas condiciones.

Una vez obtenidos todos los estadísticos de contraste (valor de α fijado al 5% en cada comparación) y las decisiones a partir de la comparación de los intervalos de confianza se ha obtenido el porcentaje de decisiones incorrectas para cada condición en cada situación estudiada.

RESULTADOS

En la figura 2 se muestran los porcentajes de decisiones incorrectas para la situación de no diferencias de medias poblacionales entre los dos grupos a comparar en tanto que en la figura 3 se muestran estos valores en la situación de si diferencias de medias poblacionales. El primer aspecto que destaca tanto en un caso como en otro es el elevado porcentaje de pares en que no se cumple la condición de aplicación de la t de *student* de grupos independientes, especialmente en los casos en que existía asimetría en uno de los dos grupos.

Por otra parte, en la situación de igualdad de medias poblacionales, se observa que en el caso de muestra de 30 o 50 el porcentaje de decisiones erróneas es muy elevado en la t de *Student* o en la comparación de los intervalos de confianza de medias, es mejor en el caso de la comparación de los intervalos de confianza de medias recortadas o de medianas, pero la prueba que muestra valores más óptimos a través de las diferentes condiciones estudiadas es el estadístico t de Yuen-Welch que mantiene el porcentaje de decisiones incorrectas alrededor de la tasa nominal de error del 5% (figura 2).

Con respecto a la situación de diferencias de medias poblacionales entre los dos grupos a comparar, para muestras de 30 o de 50 todos los criterios de decisión son adecuados a excepción de la t de S*tudent* de grupos independientes en el caso de muestras de tamaño 30. En las condiciones de tamaño de muestra 10, ninguna de las comparaciones analizadas resulta óptima, aunque la mejor sería la comparación de los intervalos de confianza de medianas a partir de la estimación adaptativa kernel (figura 3).

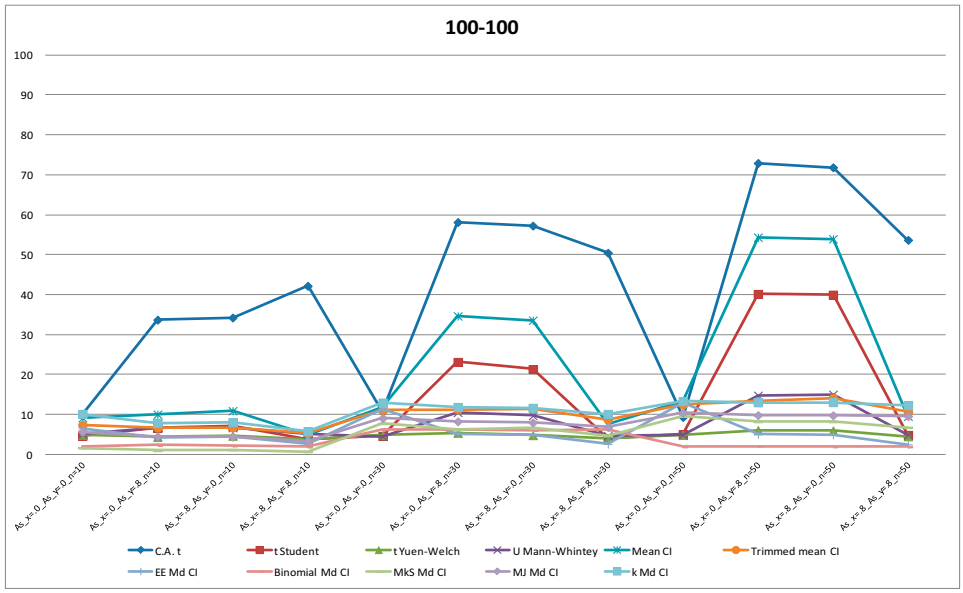


Figura 2. Porcentaje de decisiones incorrectas en la situación de igualdad de medias poblacionales en las dos muestras. (As_x: asimetría en el grupo x, As_y: asimetría en el grupo y, n: tamaño muestra, C.A. t: condición aplicación estadístico t de Student, t Student: estadístico t de Student, t Yuen Welch: estadístico t de Yuen-Welch,U Mann-Whitney: estadístico U de Mann-Whitney, Mean CI: comparación intervalos de confianza de la media, Trimmed mean CI: comparación intervalos de confianza de la media recortada, EE Md CI comparación de intervalos de confianza de la mediana según error estándar, Binomial Md CI: comparación de intervalos de confianza de la mediana según ley binomial, Mks Md CI: comparación de intervalos de confianza de la mediana según el estimador de McKean y Schraeder, MJ Md CI: comparación de intervalos de confianza de la mediana según el estimador de Marizt y Jarret y k Md CI: comparación de intervalos de confianza de la mediana según la estimación adaptativa kernel).

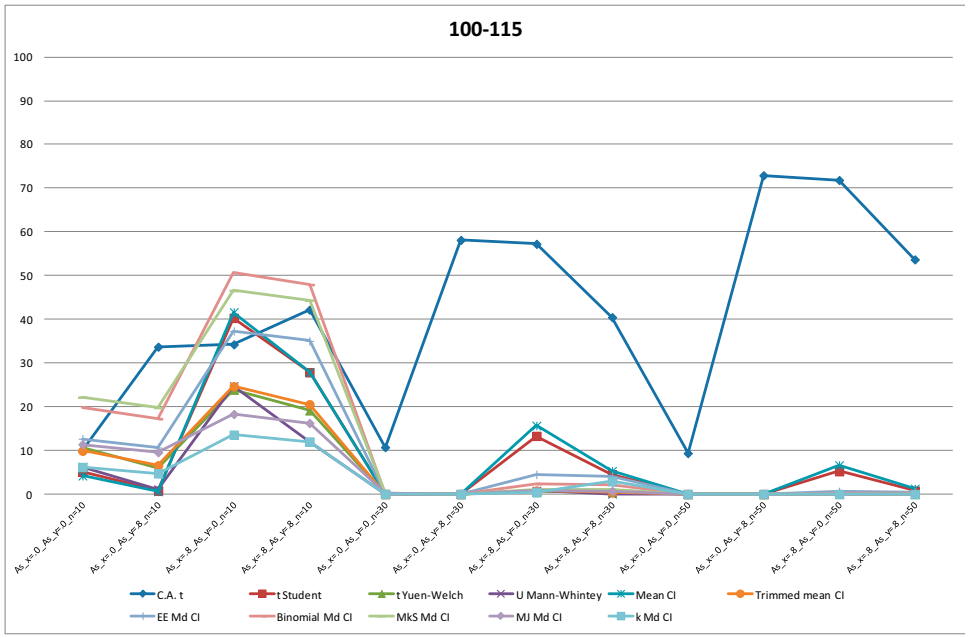


Figura 3. Porcentaje de decisiones incorrectas en la situación de desigualdad de medias poblacionales en las dos muestras (As_x: asimetría en el grupo x, As_y: asimetría en el grupo y, n: tamaño muestra, C.A. t: condición aplicación estadístico t de Student, t Student: estadístico t de Student, t Yuen Welch: estadístico t de Yuen-Welch,U Mann-Whitney: estadístico U de Mann-Whitney, Mean CI: comparación intervalos de confianza de la media, Trimmed mean CI: comparación intervalos de confianza de la media recortada, EE Md CI comparación de intervalos de confianza de la mediana según error estándar, Binomial Md CI: comparación de intervalos de confianza de la mediana según ley binomial, Mks Md CI: comparación de intervalos de confianza de la mediana según el estimador de McKean y Schraeder, MJ Md CI: comparación de intervalos de confianza de la mediana según el estimador de Marizt y Jarret y k Md CI: comparación de intervalos de confianza de la mediana según la estimación adaptativa kernel).

CONCLUSIONES

Como conclusiones más importantes del trabajo cabe comentar que la comparación de los intervalos de confianza de medianas presentan una adecuada especificidad pero no tan buena sensibilidad, en el caso de trabajar con muestras pequeñas (n =10), pero si en el caso de tamaños de muestra superiores (n = 30 o n = 50). Aspecto que también se observa en el caso de la comparación de los intervalos de confianza de medias recortadas. Aunque de todos modos es necesario explorar el uso de otros criterios de decisión en la comparación de los intervalos de confianza.

Por otra parte, cabe destacar la inadecuación del uso de pruebas clásicas como pueden ser la t de *student* de grupos independientes o la prueba no paramétrica U de Mann-Whitney. Siendo especialmente grave además del enorme porcentaje de decisiones incorrectas que se producen al utilizar estas pruebas, especialmente en el caso de la existencia de asimetría, el hecho del no cumplimiento de la igualdad de varianzas para la aplicación de la t de *student* de grupos independientes.

Para finalizar, comentar la bondad de la t de Yuen-Welch, tanto en la situación de no diferencias poblacionales como en la situación de si diferencias. De hecho, de todos los procedimientos analizados, es el que en general presenta menor tasa de error en la decisión. Posiblemente este debería ser el estadístico de elección aunque es necesario aportar más evidencias.

REFERENCIAS

APA (2001). *Publication manual of the American Psychological Association. Fitth edition*. Washington. American Psychological Association.

Bland, M. (2003). *Confidence interval for a median and other quantiles*. Documento recuperado el 5 de mayo de 2005 del Word Wide Web: http://www-users.york.ac.uk/~mb55/intro/cicent.htm.

Bonett, D.G., y Price, R.M. (2002). Statistical inference for a linear function of medians: confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, *7* (3), 370-383.

Cumming, G., y Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61* (4), 532 – 574.

Cumming, G., y Maillardet, R. (2006). Confidence intervals and replications:where will the next mean fall? *Psychological Methods*, 11 (3), 217 – 227.

DeCoster, C., y Burchill, C. (2000). *Confidence interval of the median*. Documento recuperado el 5 de mayo de 2005 del Word Wide Web: http://www.umanitoba.ca/centres/MCHP/concept/dict/ci_median.

Field, C., y Genton, M.G. (2006). The multivariate g-and-h distribution. *Technometrics*, *48* (1), 104 - 111.

Kendall, M.G. (1945). *The advanced theory of statistics. Volume I*. London: Charles Griffin & Company Limited.

Mothes, J., y Torrens-Ibern, J. (1970) *Estadística aplicada a la ingenieria* Barcelona.Ariel.

R Development Core Team. (2010). *R: A Languaje and Enviroment for Statistical Computing (Version 2.11.1) {Computer software}*. Viena: R Foundation for Statistical Computing.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, Massachussets: Addison-Wesley.

Wilcox, R.R. (2005). *Introduction to robust estimation and hipothesis testing. Second edition*. USA: Elsevier Academic Press.

Wilkinson, L., y the Task Force on Statistical Inference (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist*, *54* (8), 594 – 604.

EL ANÁLISIS TEXTUAL EN LA INVESTIGACIÓN PSICOLÓGICA. ESTADO DEL ARTE

María Jesús Carrera-Fernández, Joan Guàrdia-Olmos y Maribel Peró-Cebollero
Universidad de Barcelona
Correo electrónico: mcarrera@ub.edu

Resumen

El presente estudio es una aproximación a algunas de las problemáticas que enfrenta la investigación cualitativa en psicología y que pueden estar limitando su utilización y posterior publicación en revistas indexadas. Los autores han tomado como caso de estudio el *análisis textual*, con la finalidad de conocer su empleo en la investigación psicológica. Para ello se llevó a cabo una búsqueda bibliográfica en las bases de datos *Sciences Citation Index* (SCI), *Social Sciences Citation Index* (SSCI) y *Arts and Humanities Citation Index* (A&HCI), de la Web of Knowledge, con las palabras clave *«textual analysis»* y *«text analysis»*. Posteriormente se analizó a profundidad una muestra de los artículos encontrados, contemplando aspectos de publicación, sustantivos, metodológicos y contextuales. Los resultados encontrados llevan a los autores a plantearse que el «análisis textual» no puede considerarse una metodología específica y que el empleo de este término conlleva confusiones.

En la investigación cualitativa se emplea un gran número de términos para representar la variedad de aproximaciones y métodos (Denzin & Lincoln, 2000). Hay casos en los que el mismo nombre se emplea para referirse a procedimientos distintos, o el mismo procedimiento recibe diferentes nombres. En las publicaciones de investigación cualitativa algunos títulos, abstracts o keywords mencionan la aproximación o método usado, otros la posición filosófica del método y otros más alguna técnica de recolección de datos (Marchel & Owens, 2007), lo que dificulta cualquier intento de revisión bibliográfica.

Los primeros análisis sistemáticos de textos se remontan a las persecuciones inquisitoriales realizadas en el siglo 17 (Krippendorff, 2004a). Si se intenta seguir el desarrollo histórico del análisis de textos se descubre que en algún momento se comienza a hablar del análisis de contenido. En 1903 Eugen Löbl publicó un esquema de clasificación para analizar el contenido de los periódicos (Krippendorff, 2004a). El inicio del análisis de contenido era analizar textos, por lo tanto podría haberse llamado análisis de contenido o análisis de textos.

Hewson (2008) hace referencia al análisis textual como una técnica que emplea procedimientos cualitativos para evaluar la importancia de ideas o significados dentro de un documento; remitiendo para mayor discusión a Scott (2006). Sin embargo, la referencia de Scott se titula *«Content analysis»*, lo que retorna nuevamente a la confusión inicial. Contribuyendo a esta confusión encontramos que el concepto «text analysis» es empleado por autores como Popping (2007) que sostiene que en el pasado los científicos sociales desarrollaban análisis de contenido, pero que hoy en día se usa el término thematic text analysis.

En los manuales de investigación cualitativa más reconocidos el análisis textual no se incluye, sin embargo el término se emplea en foros, congresos y publicaciones. Ante esto nos preguntamos si realmente puede ser considerado un método, buscando profundizar en los siguientes interrogantes: ¿Existen publicaciones psicológicas que empleen el análisis textual? ¿Es el análisis textual un método dentro de la investigación cualitativa en psicología?

MÉTODO

Llevamos a cabo una búsqueda bibliográfica en las bases de datos SCI, SSCI y A&HCI, de la Web of Knowledge. Las palabras clave fueron *«textual analysis»* y *«text analysis»* en el *título*, el tipo de documento fue *artículos*, sin fecha límite de publicación y de cualquier área de la psicología. Seleccionamos una muestra de los artículos que fue analizada a profundidad, incluyendo aspectos de publicación, sustantivos, metodológicos y contextuales.

RESULTADOS

La búsqueda dio un resultado de 25 artículos. Se realizó el mes de Junio de 2010 y se actualizó periódicamente hasta Noviembre del mismo año. Del total de artículos se tomó una muestra de 15 (tabla 1), el criterio de inclusión fue la disponibilidad del texto completo.

Tabla 1. Artículos que formaron parte del estudio

Artículo	Revista
Construct Validation Using Computer-Aided Text Analysis (CATA) An Illustration Using Entrepreneurial Orientation	Organizational Research Methods
Religious Fundamentalism and Responses to Mortality Salience: A Quantitative Text Analysis	International J. for the Psychology of Religion
Therapeutic factors and language patterns in group therapy application of computer-assisted text analysis to the examination of microprocesses in group therapy: Preliminary findings	Psychotherapy Research
Truly toffee and raisin hell: A textual analysis of lipstick names	Sex Roles
Analyzing Ghanaian emotions through narrative: a textual analysis of Ama Ata Aidoo's Novel Changes	Journal of Black Psychology
Evaluation of computerized text analysis in an Internet breast cancer support group	Computers in Human Behavior
Comparison of textual analysis applied to two lectures written three years apart by one author: The language satellites	Psychological Reports
The assessment of cross-cultural experience: measuring awareness through critical text analysis	International Journal of Intercultural Relations
ETAT: Expository Text Analysis Tool	Behavior Research Methods, Instruments & Computers.
The relationship among attachment representation, emotion-abstraction patterns and narrative style: a computer-based text analysis of the adult attachment interview	Psychotherapy Research
Textual analysis of addictive behavior of cigarettes' smokers undergoing stop smoking treatment	Revue Europeenne de Psychologie Appliquee
Extracting team mental models through textual analysis	J. Organizational Behavior
Hazardous measures: an interpretative textual analysis of quantitative sensemaking during crisis	J. Organizational Behavior
Decentering therapy: Textual analysis of a narrative therapy session	Family Process
Protocol modeling, textual analysis, the bifurcation bootstrapping method and Convince me - computer-based techniques for studying beliefs and their revision	Behavior Research Methods, Instruments & Computers.

ASPECTOS DE PUBLICACIÓN Y CONTEXTUALES

Respecto al número de autores son menos las publicaciones en solitario (4; 26.66%) y más en pareja (6; 40%) o equipo (5; 33.33%). Los factores de impacto más altos tienden a provenir de las áreas de *Applied, Management* y *Clinical,* seguidas por *Business* (tabla 2).

El país con el mayor número de publicaciones es Estados Unidos con 6 (40%), seguido por Alemania y Francia con 2 cada uno (13.33%) y por Bélgica, Canadá y Japón con 1 (6.66%). El idioma de publicación mayoritario fue el inglés (93.33%), excepto un artículo publicado en francés. En los corpus la predominancia del inglés se mantuvo, con nueve (60%) corpus en ese idioma.

Tabla 2. Aspectos de publicación

Artículo	Año de publicación	Número de autores	Género del 1er autor	Citas	Factor de impacto (JCR)	Área (ISI)
1	2010	4	Masculino	0	3.019	Applied; Management
2	2008	1	Masculino	1	_	Multidisciplinary; Religion
3	2008	2	Femenino	0	1.579	Clinical
4	2007	1	Femenino	0	0.652	Developmental; Social; Women's Studies
5	2007	2	Masculino	0	0.860	Multidisciplinary
6	2005	7	Masculino	13	1.116	Multidisciplinary; Experimental
7	2003	2	Femenino	3	0.277	Multidisciplinary
8	2002	5	Masculino	1	0.430	Social; Social Sciences; Sociology
9	2002	6	Masculino	4	0.851	Mathematical; Experimental
10	2000	2	Femenino	28	1.439	Clinical
11	1999	3	Femenino	1	_	Applied
12	1997	1	Femenino	48	1.990	Business; Applied; Management
13	1997	1	Masculino	21	1.990	Business; Applied; Management
14	1997	2	Masculino	30	1.275	Clinical; Family Studies
15	1995	2	Masculino	6	_	Mathematical; Experimental

Nota. El JCR corresponde al año de publicación, excepto los que se muestran en cursivas que corresponden al año 2009.

ASPECTOS SUSTANTIVOS

Sólo 11 de los 15 artículos analizados especifican su objetivo. El tipo de artículo se catalogó como aplicado o no aplicado, siendo sólo dos artículos no aplicados.

Es notorio que temas tan diferentes llevan a corpus también muy variados en temática, extensión y forma de obtención (tabla 3). Cada artículo se clasificó entre una y tres áreas (tabla 2), incluyéndose 14 áreas pertenecientes a la psicología o estrechamente vinculadas a ella.

Tabla 3. Aspectos sustantivos

Participantes/Muestra	Corpus
950 firmas	655 cartas de CEOs (1,512 páginas)
400 personas	400 textos sobre la muerte o el dolor dental
5 mujeres	Transcripciones de 42 sesiones de grupo
52 marcas	1722 nombres de pintalabios
1 novela	«Changes»
9 mujeres	521 posts en un forum
2 ensayos	«Double talk» y «Thanatol»
176 estudiantes	176 reportes de experiencia práctica
40 padres	Transcripciones de la Adult Attachment Interview
50 fumadores	Transcripciones de 50 entrevistas no directivas
41 estudiantes	Respuestas a dos preguntas abiertas
1 investigación pública	217 páginas del procedimiento oficial y 16 páginas del reporte final
1 sesión de terapia	Videograbación de una terapia de pareja

ASPECTOS METODOLÓGICOS

De acuerdo al esquema de Madill y Gough (2008), los métodos de recolección de datos más utilizados fueron los *naturales* (6; 46.16%), la *entrevista* (4; 30.77%) y los *estructurados* (3; 23.08%).

Respecto a los métodos de análisis sólo cuatro artículos hacían referencia al análisis textual pero siempre con puntualizaciones: análisis textual cualitativo; análisis textual, mejor descrito como híbrido de análisis de la retórica y estudio de caso instrumental; análisis textual interpretativo; y características de análisis del discurso, conversacional, textual y de narrativas. Encontramos que los más comunes podrían ser considerados *estructurados* (5; 38.46%), *temáticos* (3; 23.08%) y *discursivos* (3; 23.08%) (Madill & Gough, 2008). Hubo también un artículo en el que no se especifica cómo se realizó el análisis. Siete de los 15 artículos (46.66%) establecieron hipótesis.

Un aspecto que decidimos incluir posteriormente fue el análisis estadístico. Sólo dos de los artículos no utilizaron ninguna prueba estadística. Los más utilizados fueron los *test de relación* (ocho artículos), las *correlaciones* (cinco), los esta-

dísticos *descriptivos* (cuatro), los *multivariados* (cuatro) y los *análisis de varianza* (tres).

De los artículos estudiados 11 (73.33%) utilizaron algún software. Los programas LIWC, TAS/C y ALCESTE fueron utilizados cada uno en dos artículos y los programas DICTION, ETAT y TACT en uno. En seis casos tanto el corpus como el software con el cual se analizó estaban en inglés; en dos casos ambos estaban en francés y en un caso en alemán. Hay un artículo del que se desconoce el idioma del corpus; además de otro en el que no concuerda el idioma del corpus con el de los diccionarios, mencionándose que se adaptaron pero sin aclarar el procedimiento de traducción.

Finalmente juzgamos importante conocer la lógica que se había seguido en el análisis. Consideramos que era inductiva cuando las categorías y/o los diccionarios eran creados basados en la propia muestra, es decir *text-based*; mientras que era deductiva cuando se desarrollaban con independencia de cualquier texto particular, *code-based*. Siete de los 15 artículos (46.66%) se consideraron inductivos, cinco (33.33%) deductivos y los tres restantes (20%) mixtos.

CONCLUSIONES Y DISCUSIÓN

Se encontraron muy pocos artículos con las palabras clave en el título (sólo 25 desde 1980) en comparación con otros métodos con los mismos criterios de búsqueda (pe., «content analysis» da un resultado de 455 artículos). La mayoría de los autores definían en detalle los pasos seguidos, pero no especificaban qué método era o hacían una mezcla de varios métodos. Ninguno explicitaba que el método era análisis textual. La combinación de métodos podría considerarse enriquecedora, es como tener un objeto de estudio e ir construyendo el método más adecuado para aproximarse a él. Sin embargo, puede llegar a ocasionar confusión e impactar en la validez y la replicabilidad.

En general, el análisis de datos era cuantitativo más que cualitativo. Varios artículos empezaban con una clasificación y categorización para posteriormente pasar a un análisis cuantitativo. ¿Puede esto considerarse perjudicial para la investigación cualitativa? En absoluto, siempre y cuando se tenga claro qué es lo que se está trabajando y se especifique el método seguido. Tanto la investigación cualitativa como la cuantitativa son propensas a adquirir características del otro modelo lógico. Lo que debe tenerse presente es que emplear un modelo hipotético-deductivo en una investigación cualitativa puede ocasionar que se pasen por alto datos relevantes, es decir, que el investigador se predisponga a considerar como dato aspectos que concuerden con alguna teoría en particular y no otros. Un ejemplo de lo anterior es el trabajo con categorías *code-based*.

Al trabajar con textos el idioma cobra una especial relevancia y la traducción se vuelve un tema delicado. La traducción de un corpus o de los diccionarios debería ser evitada o al menos ser especificada y seguir un procedimiento cuidadoso.

Respecto al software debe tenerse presente que su papel es el de una herramienta más. Desafortunadamente se observó que algunos autores para definir el método describían cómo funcionaba el software que empleaban, dándole una prioridad que no le corresponde.

Este estudio tiene varias limitaciones. Una de ellas es el alcance de las bases de datos, además de los términos de búsqueda, específicamente que las palabras clave debían aparecer en el título. Hemos considerado adecuado hacerlo así buscando que los artículos fueran de análisis textual y no apareciera sólo de manera tangencial. Otra limitación es el escaso número de artículos analizados. A pesar de ello, el análisis nos proporciona evidencias para responder los interrogantes planteados. Hemos visto que sí existen publicaciones que emplean el análisis textual, o tal vez sería más certero decir que emplean algún procedimiento al que llaman análisis textual; pero no hemos podido encontrar un método específico al que correspondan. Es posible decir que el análisis textual no es un método particular, sino un concepto que se emplea para referirse a aproximaciones distintas o para hacer alusión a que se trabaja con textos, por lo que no puede considerarse un método en sí mismo.

Consideramos importante concluir señalando que es necesaria la generación de un lenguaje común en la investigación cualitativa. Los términos genéricos, como en este caso el análisis textual, deberían ser evitados. Hoy en día se están haciendo interesantes aproximaciones a los textos en la psicología (véase por ejemplo a Mehl & Gill, 2010) que están incorporando aspectos más cualitativos. Por nuestra parte creemos que el análisis de los textos aún tiene mucho que aportar a la psicología.

NOTA DE LOS AUTORES

Esta investigación ha recibido financiación del *Comissionat per a Universitats i Recerca* de la *Generalitat de Catalunya* y el European Social Fund.

REFERENCIAS

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.) Thousand Oaks, CA: Sage.

Hewson, C. (2008). Internet-mediated research as an emergent method and its potencial role in facilitating mixed methods research. In S. N. Hesse-Biber & P. Leavy (Eds.) *Handbook of emergent methods* (pp. 543-570). Nueva York: Guilford Press.

Krippendorff, K. (2004a). *Content Analysis. An introduction to its methodology*. 2nd ed. Thousand Oaks, CA: Sage.

Krippendorff, K. (2004b). Measuring the reliability of qualitative text analysis data. *Quality & Quantity, 38*, 787-800.

Madill, A., & Gough, B. (2008). Qualitative research and its place in psychological science. *Psychological Methods, 13*, 254-271. doi: 10.1037/a0013220

Marchel, C., & Owens, S. (2007). Qualitative research in psychology: could William James get a job? *History of Psychology, 10* (4), 301-324. doi: 10.1037/1093-4510.10.4.301

Mehl, M. R., & Gill, A. J. (2010). Automatic text analysis. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 109-127). Washington, DC: American Psychological Association.

Popping, R. (2007). Text analysis for knowledge graphs. *Quality & Quantity, 41*, 691-709. doi: 10.1007/s11135-006-9020-z

Scott, J. (2006). Content analysis. In V. Jupp (Ed.), *The Sage dictionary of social research methods* (pp. 40-41). London: Sage.

EXPLORATORY STRUCTURAL EQUATION MODEL (ESEM): CONCEPT, OBJECTIVES AND OPERATIVIZATION. APPLICATION TO THE UNIVERSITY STUDENTS' EVALUATION

Joan Guàrdia, Maribel Peró y Sonia Benítez
Universidad de Barcelona
Correo electrónico: jguardia@ub.edu

Abstract

The aim is to approach the development of the recently appeared Exploratory Structural Equation Models (ESEM) as a complement and alternative to the Confirmatory Factorial Analysis derived from the models of measurement of the Structural Equation Models (SEM). That alternative is based on the new conception of exploratory factorial load matrix and it has yielded some evidence of appropriateness for the determination of factorial structures of measurement in multilevel designs. However, its nonexistent tradition, given how recent the approach is, prevents us from having evidence as to how that algorithm can develop in wider computer systems (such as R), or how its development may be influenced by the elements derived from the distribution of the observed variables (study of asymmetry and kurtosis), of the sample size (for example, with different-size samples), or among others, of the impact of the type of rotation formulated (varimax, oblimin, quartimin, quartimax, geomin, promax, etc.). Moreover, neither is there evidence of the effect and possible application of this type of statistical technology in multigroup designs with large samples and with the limits of applied research. We know how it works with large samples in multilevel designs, but we have no information about that effect on other types of designs. We will analyze the possible goodness of applying the ESEM in the factorialization of psychological phenomena. In fact, we show several result obtained from the application to ESEM algorithm from the results of a sample of university students in relation with their exam response.

Los Modelos de Ecuaciones Estructurales son una técnica que ha sido utilizada extensamente en la estimación de estructuras factoriales y, por tanto, en estudios psicométricos de fenómenos psicológicos complejos.

Tal y como apuntan Asparouhov y Muthén (2009) en su estudio, una de las ventajas de este tipo de técnica es que el uso del análisis factorial confirmatorio (CFA) en el modelo de medida del SEM permite al investigador proponer una estructura simple del modelo de medida debido a que el investigador incorpora el conocimiento sustantivo previo en forma de determinadas restricciones en el modelo de medida, lo que da lugar a que la definición de la variable latente tenga una mejor fundamentación teó-

rica; y esto permite que los modelos propuestos sean más parsimoniosos. Sin embargo, al proponer una estructura simple donde cada ítem carga sólo en un factor y el resto de cargas factoriales cruzadas son fijadas a 0, puede obligar a que el investigador esté especificando un modelo más parsimonioso de lo que sería adecuado para los datos, y en ocasiones eso podría dar lugar a la obtención de índices de bondad de ajuste no del todo adecuados (Asparouhov y Muthén, 2009). En esta misma línea, el mismo estudio concluye que la restricción en CFA de especificar a 0 aquellas cargas cruzadas bajas puede dar lugar a que se sobreestimen los coeficientes de correlación entre los factores y, por tanto, las relaciones estructurales estén distorsionadas.

En este sentido, dado que el uso del CFA en el modelo de medida de los SEM puede haber contribuido a que se ponga en duda la credibilidad y replicabilidad del modelo propuesto en determinados ámbitos de aplicación de esta técnica (Asparouhov y Muthén, 2009), recientemente a aparecido una nueva aproximación a los SEM, los Modelos de Ecuaciones Estructurales Exploratorios (ESEM). Esta alternativa exploratoria de los SEM se basa en la nueva concepción de la matriz de carga factorial exploratoria, ya que se pueden utilizar partes del modelo mediante EFA con matrices de cargas factoriales rotadas además de o en lugar de partes del modelo de medida mediante CFA. Es decir, a diferencia de lo que ocurre en el CFA, el modelo de medida mediante EFA no requiere estrictas restricciones de las cargas factoriales cruzadas. Con el añadido de que, igual que en el clásico enfoque SEM, da acceso a los usuales parámetros como, por ejemplo, correlaciones residuales, regresiones de los factores en relación, y correlaciones entre los factores.

Marsh, Muthén, Asparaouhov, Lüdtke, Robitzsch, Morin, y Trautwein (2009) llevaron a cabo una primera aplicación de los ESEM a las valoraciones de los estudiantes sobre la docencia universitaria, demostrando el poder y la flexibilidad de esta reciente aproximación exploratoria de los modelos estructurales, integrando las ventajas de los SEM, el CFA y el EFA. En esta misma línea, Marsh, Muthén, Morin, Lüdtke, Asparaouhov, Trautwein, y Nagengast (2010) presentaron ESEM como una posible alternativa para el análisis de la estructura factorial del Modelo de personalidad *Big Five.*

Por tanto, el objetivo del presente estudio es analizar la posible bondad de ajuste de una aplicación de ESEM en la factorialización de fenómenos psicológicos, concretamente en nuestro caso, de la evaluación de los estudiantes sobre la docencia universitaria, realizando una comparación entre la nueva aproximación y el clásico enfoque SEM.

MÉTODO

Participantes

Para llevar a cabo el presente estudio se utilizó una muestra final de 662 alumnos, matriculados en la asignatura *Análisis de datos en Psicología* durante el primer semestre del curso académico 2008-2009.

Instrumento

Se utilizó un cuestionado elaborado para este estudio, consistente en 30 ítems agrupados en 6 tipos de materiales o recursos docentes diferentes utilizados en la asignatura: Bibliografía recomendada, Manual de referencia, Diapositivas, Dosier de prácticas, Actividades de teoría y Actividades prácticas. Cada uno de estos seis materiales estaba definido por los mismos cinco adjetivos, operacionalidados en un continum entre dos polos semánticamente opuestos (Útil – No útil, Facilidad – Dificultad, Organizado – No organizado, Eficaz –No eficaz, Interesante – No interesante), que los estudiantes debían valorar respecto a cada recurso docente en una escala continua de 10 cm, correspondiendo el 0 al polo negativo del adjetivo y el 10 al polo positivo del mismo.

Análisis de datos

El análisis de los datos llevado a cabo consistió, en primer lugar, en un análisis SEM mediante el método de rotación Oblimin (asumiendo correlación entre los factores), con el total de la muestra para probar la estructura factorial del cuestionario mediante este enfoque clásico. Posteriormente, se realizó el mismo análisis a través de un enfoque estructural exploratorio (ESEM), para comparar la bondad de ajuste entre las dos aproximaciones. Y, finalmente, se llevó a cabo un análisis multigrupo mediante ESEM con el objetivo de probar la invariancia de la estructura factorial del instrumento a través de cuatro métodos docentes diferentes (correspondientes a cuatro profesores que imparten la misma asignatura).

Resultados

A continuación se describirán los resultados obtenidos de los correspondientes análisis.

En primer lugar, a partir del análisis SEM se obtuvo una estructura factorial del cuestionario analizado de 7 factores, es decir, a demás de los seis factores correspondientes a los diferentes tipos de materiales docentes utilizados en la asignatura, aparece un séptimo factor formado por tres ítems (interés/no interés de la bibliografía recomendada, y facilidad/dificultad y interés/no interés del manual de referencia) que pertenecen también a alguno de los otros seis factores. Las cargas factoriales de los ítems que aparecen en la matriz factorial rotada de esta solución oscilan entre .612 y .947, siendo entre .416 y .689 en aquellos ítems que, según este enfoque clásico, pertenecen a dos factores. Así mismo, los coeficientes de correlación entre factores oscilan entre .150 y .389, siendo estadísticamente significativos en todos los casos.

Posteriormente, se realizó el mismo análisis mediante ESEM. En este análisis se obtuvo una estructura del cuestionario de 6 factores, en la línea de la estructura original del mismo. En este caso, las cargas factoriales de los ítems que aparecen

en la matriz factorial rotada oscilan entre .718 y .944, de forma similar que en SEM. En cambio, sí que se observan diferencias en los coeficientes de correlación entre factores, ya que mediante ESEM éstos tienden a ser más elevados, oscilando entre .219 y .461 (todos estadísticamente significativos).

Por último, respecto al análisis multigrupo mediante ESEM para probar la invariancia de la estructura factorial de los diferentes materiales docentes utilizados a través de 4 docentes diferentes, en la tabla 1 se presenta un resumen de los índices de bondad de ajuste.

Tabla 1. Resumen de los índices de bondad de ajuste del análisis multigrupo

	Modelo ESEM (4 grupos)	c2/g.l.	GFI	CFI	BBNFI	BBNNFI	RMSEA	R2	α
1	IN = None	1456.22 / 238	.946	.936	.942	.929	.07	.86	.94
2	IN = FL	1456.22 / 283	.944	.933	.938	.931	.07	.81	.95
3	IN = FL, Unq	1456.22 / 328	.944	.935	.937	.930	.06	.93	.97
4	IN = FL, FVCV	1456.22 / 336	.945	.937	.936	.932	.08	.84	.96
5	IN = FL, INT	1456.22 / 412	.944	.928	.935	.933	.07	.79	.95
6	IN = FL, Unq, FVCV	1456.22 / 418	.941	.929	.936	.926	.08	.80	.94
7	IN = FL, Unq, INT	1456.22 / 461	.940	.930	.940	.927	.06	.81	.96
8	IN = FL, FVCV, INT	1456.22 / 398	.939	.931	.944	.930	.06	.82	.97
9	IN = FL, Unq, FVCV, INT	1456.22 / 451	.939	.930	.944	.933	.06	.81	.94
10	IN = FL, INT, FMn	1456.22 / 461	.941	.929	.945	.934	.09	.82	.91
11	IN = FL, Unq, INT, FMn	1456.22 / 459	.940	.933	.949	.935	.06	.84	.92
12	IN = FL, FVCV, INT, FMn	1456.22 / 429	.942	.932	.938	.926	.08	.81	.95
13	IN = FL, Unq, FVCV, INT, FMn	1456.22 / 438	.943	.935	.934	.922	.07	.85	.94
	SEM Total muestra	2962.91 / 390	.857	.837	.817	.818	.0923	.72	.917
	ESEM Total muestra	1456.22 / 268	.942	.933	.938	.933	.0732	.83	.952

Nota: *GFI:* Índice de ajuste LISREL; *CFI:* Índice de ajuste comparativo; *BBNFI:* Bentler-Bonett Normado; *BBNNFI:* Bentler-Bonet no normado; *RMSEA:* Residual medio cuadrado de aproximación; *IN:* conjunto de parámetros con la restricción de permanecer invariantes a través de los múltiples grupos; *FL:* Cargas factoriales; *Unq:* Unicidad del ítem; *FVCV:* Variancias-covariacias de los factores; *INT:* Interceptos de los ítems; *FMn:* Medias de los factores.

Para este análisis la muestra estaba clasificada en 4 grupos correspondientes a cuatro profesores de la asignatura y, por tanto, a cuatro estrategias docentes diferentes. Así mismo, utilizamos la taxonomía propuesta por Marsh et al. (2009), que operacionalizaron a través de 13 modelos parciales, siendo el primero de ellos el menos restrictivo (sin ninguna restricción de invariancia en la estimación de parámetros) hasta llegar al último modelo con mayor restricción (modelo de completa invariancia en la estimación de los parámetros a través de los grupos). En este análisis, se obtuvieron unos índices de bondad de ajuste muy similares en los 13

modelos y cuyos valores pueden considerarse bastante aceptables, lo que indicaría que el modelo propuesto ajusta bien a los datos. Además, en el análisis factorial con el total de la muestra observamos que los índices de ajuste mediante SEM no son del todo aceptables, mientras que en ESEM hemos obtenido buen ajuste del modelo factorial y mayor varianza explicada.

DISCUSIÓN Y CONCLUSIONES

A partir de los resultados presentados anteriormente podemos concluir, en la línea de lo que plantearon Marsh et al. (2009), que el uso de una aproximación exploratoria de los modelos estructurales (ESEM) en el análisis de estructuras factoriales proporciona mejores índices de bondad de ajuste en comparación con el clásico SEM y, además, en nuestro caso eso indica que se confirmaría la estructura factorial originaria del instrumento.

Sin embargo, cabe apuntar que se ha observado mayor correlación entre los factores en el análisis ESEM, resultados que irían en contra de lo aportado por Asparouhov y Muthén (2009), quienes concluyeron que mediante un enfoque SEM se obtienen correlaciones entre factores sobreestimadas y, por tanto, relaciones estructurales distorsionadas, debido a que en el modelo de medida del CFA propio del clásico enfoque SEM se fuerza a que cada indicador forme parte sólo de un factor y por tanto el resto de cargas factoriales se fijan a 0. Nuestro estudio aporta resultados contrarios, lo que podría indicarnos una menor fiabilidad del modelo ESEM.

Por otro lado, en el análisis multigrupo mediante ESEM obtuvimos unos índices de bondad de ajuste aceptables en los 13 modelos parciales, y no presentaron grandes diferencias entre ellos. Por tanto, estos resultados nos hacen pensar en la robustez y estabilidad de la estructura factorial del cuestionario a través de los cuatro grupos de docencia diferentes.

Por tanto, a modo de conclusión, podemos decir que los resultados del presente estudio inducen a pensar que los Modelos de Ecuaciones estructurales Exploratorios (ESEM) pueden ser un buen complemento y alternativa al CFA derivado del modelo de medida de los SEM. De todas formas, se trata de una primera aproximación ya que es un enfoque muy reciente. Por lo que las investigaciones futuras deben ir encaminadas a aportar nuevas evidencias en esta línea, así como a analizar el efecto del método de rotación utilizado (varimax, oblimin, quartimin, quartimax, geomin, etc.) para el análisis de estructuras factoriales en fenómenos psicológicos.

NOTA DE LOS AUTORES

El estudio ha contado con el apoyo económico de la *Secretaria d'Universitats i Investigació (SUR) del Departament d'Economia i Coneixement (DEC) de Generalitat de Catalunya de Catalunya* i del *Fons Social Europeu (FSE)*.

REFERENCIAS

Asparouhov, T. y Muthén, B. (2009). *Exploratory structural equation modeling. Structural Equation Modeling, 16* (3), 397–438.

Bollen, K. A. y Davis, W. R. (2009). Two rules of identification for structural equation models. *Structural Equation Modeling, 16* (3), 523–536.

Marsh, H.W., Muthén, B., Asparaouhov, T., Lüdtke, O., Robitzsch, A., Morin, A., y Trautwein, U. (2009). Exploratory structural equation modeling, interpretating CFA and EFA: Application to student's evaluations of university teaching. *Structural Equation Modeling, 16,* 439–476.

Marsh, H.W., Muthén, B., Morin, A., Lüdtke, O., Asparaouhov, T., Trautwein, U., y Nagengast, B. (2010). A new look at the Big Five Factor structure through Exploratory Structural Equation Modeling. *Psychological Assessment, 22,* (3), 471–491.

APLICACIONES ACTUALES Y NUEVAS PERSPECTIVAS DE LOS MODELOS DE ANÁLISIS MULTINIVEL EN CIENCIAS SOCIALES Y DE LA SALUD

Coordinadoras: Nekane Balluerka y Arantxa Gorostiaga
Universidad del País Vasco

Los modelos de análisis multinivel brindan excelentes posibilidades para examinar las relaciones existentes entre los individuos y los contextos en los que éstos se ubican, por lo que se han convertido en una de las aproximaciones más utilizadas para el modelado de datos en Ciencias Sociales y de la Salud. Desde la perspectiva transversal, posibilitan explicar fenómenos o relaciones que se producen a nivel individual tomando en consideración variables de nivel grupal u organizacional permitiendo representar más adecuadamente la realidad, mientras que desde el enfoque longitudinal brindan la posibilidad de modelar el cambio que se produce en los sujetos a lo largo del tiempo asociándolo con la pertenencia a determinados grupos o contextos. El presente simposio pretende ser un punto de encuentro o de intercambio de experiencias en las que se aplican los modelos de análisis multinivel para el modelado de datos tanto transversales como longitudinales en los ámbitos educativo y organizacional. Para ello, se presentan 3 trabajos. En el primero, Balluerka y cols. examinan la relación existente entre la Inteligencia Emocional Individual, la Inteligencia Emocional Grupal y la dificultad para experimentar alegría o felicidad en una muestra de estudiantes de ESO y de Bachillerato. En el segundo, Elorza y cols. analizan la influencia que ejerce la percepción de los trabajadores sobre el sistema de prácticas de gestión de personas en el rendimiento organizativo. Estos dos trabajos, aunque están ubicados en distintos ámbitos, se desarrollan desde un enfoque transversal. El último, por el contrario, modela datos de carácter longitudinal. Así, McArdle presenta las ventajas de los modelos de ecuaciones estructurales, en concreto, del modelado multinivel y del modelado de clase latente para el análisis de las curvas de crecimiento latente partiendo de un conjunto de datos que representan el desarrollo de las habilidades intelectuales a lo largo del ciclo vital.

PALABRAS CLAVE: Modelos de regresión multinivel, Curvas de crecimiento latente, Modelos de ecuaciones estructurales.

INFLUENCE OF INDIVIDUAL AND GROUP EMOTIONAL INTELLIGENCE ON DEPRESSION: A MULTI-LEVEL APPROACH

Nekane Balluerka, Arantxa Gorostiaga, Aitor Aritzeta, Itziar Alonso-Arbiol and Mikel Haranburu
Universidad del País Vasco
E-mail: nekane.balluerka@ehu.es

Abstract

In the present study we examine, from a multilevel approach, the relationship between Individual Emotional Intelligence, Group Emotional Intelligence and Depression in a sample of 2,182 adolescents (1,127 female and 1,055 male) aged between 12 and 18 (M = 14.51, SD = 1.55). They attended 14 secondary schools in the Basque Country and were grouped into 118 different classrooms. A two-level model (students nested in classrooms) with four predictor variables of level 1 (sex, attention, understanding and regulation of emotions) and one variable of level 2 (emotional intelligence in the classroom) was used to examine the influence they have on depression in adolescence. The results indicate that increasing both understanding and the ability to regulate emotions at individual level reduces adolescents' depression. The group emotional intelligence reduces the average variability observed between the classes in the level of depression and has a negative relationship with said construct. However, this group-level co-variable does not explain the observed differences between classrooms with regard to the relationship between emotion regulation and depression.

Literature has broadly related self-perceived EI and psychological adjustment. In this regard, individuals with low emotional clarity and a low ability to regulate their own emotional states show poor emotional adjustment (Salovey, 2001; Fernández-Berrocal, Salovey, Vera, Extremera & Ramos, 2005). High self-perceived EI has been shown to be related, among adolescent populations, to lower levels of violence, perceived stress and depressive thoughts (Downey, Johnston, Hansen, Birney & Stough, 2010). Furthermore, depressed individuals score lower than non-depressed individuals on clarity of emotions and emotional repair (Fernández-Berrocal, Alcaide, Extremera & Pizarro, 2006).

On the other hand, several studies have linked gender and depression. It has been found that women are more likely to ruminate on negative feelings (Nolen-Hoeksema & Girgus, 1994) and such rumination is associated with higher levels of depression (Abela, Brozina & Haigh, 2002).

In the educational field, students can be considered as members of the classroom, which is a higher relevant meaningful unit of shared emotional experience. The classroom can be considered as an emotional reference group in which individual-level affective experiences combine to form group affect via emotional contagion (Hatfield, Caccioppo & Rapson, 1994). The classroom allows a rich combination of information processing and emotional responding that influences students' learning processes and coping behaviours (Meyer & Turner, 2006). Research in the field of flow theory (Csikszentmihalyi & Csikszentmihalyi, 1988) has found that students in high involvement classrooms (defined by happiness, motivation to learn and students' identification with the classroom) report significantly more experiences of flow than students belonging to low involvement classrooms (Turner, Meyer, Cox, Logan, DiCintio & Thomas, 1998).

Although the adolescents' perception of their emotional abilities may vary depending on the characteristics of the group in which they are immersed (Skinner & Zimmer-Gembeck, 2007), little research has been conducted to examine the interaction between classroom EI and individual attitudes and behaviour. Trying to shed some light on this topic, the main aim of our study is to examine whether a higher level of Individual self-perceived EI reduces depressive feelings in adolescence and to analyse if such a relationship is context-dependent, taking classroom EI as a higher-level influencing context. A secondary aim of the research is to examine whether adolescent women show higher levels of depression than men.

METHOD

Participants and Procedure

The sample was made up of 2,182 adolescents (1,127 female and 1,055 male students) aged between 12 and 18 (M = 14.51; SD = 1.55), all attending secondary schools in the Basque Country (northern Spain). The data collection was carried out in classrooms during normal school days by two researchers. The study followed the ethical guidelines of the Spanish Official Association of Psychologists and had the approval of the Ethical Committee for Research on Humans of the University of the Basque Country.

Instruments

Short Version of the Trait Meta-Mood Scale for adolescents (TMMS-23; Salguero, Fernández-Berrocal, Balluerka & Aritzeta, 2010; in its Basque Version Gorostiaga, Balluerka, Aritzeta, Haranburu & Alonso-Arbiol, 2011). Through three subscales, the TMMS-23 is a self-report tool that assesses the extent to which people: a) pay attention to and value their feelings, b) feel clear rather than confused about their feelings, and c) use positive thinking to repair negative moods. It

includes 23 items to be answered on a 5-point Likert scale. The tool has shown good psychometric properties in various studies (Gorostiaga et al., 2011).

Basque Group Trait Meta-Mood Scale (G-TMMS; Aritzeta, Balluerka, Alonso-Arbiol, Haranburu, Gorostiaga & Gartzia, in progress). The G-TMMS is aimed to measure perceived emotional intelligence at group level. It assesses the extent to which people attend to and value the feelings of the group, feel clear rather than confused about such feelings and use positive thinking to repair negative group moods. It includes 16 items to be answered on a 5-point Likert scale. The G-TMMS has shown adequate reliability and validity indices in a population of classrooms at secondary school level (Aritzeta et al., in progress).

Children's Depression Scale (CDS; Lang & Tisher, 1978; in its Basque version, Balluerka, Gorostiaga, & Haranburu, 2012). The CDS is a self-report tool that assesses depression in children and adolescents aged between 8 and 16 years. Its short version consists of 37 items to be answered on a 5 point Likert scale which assess depressive and Positive Dimensions. In this study only the positive dimension was used. The psychometric properties of the Basque CDS have been shown to be appropriate (Balluerka et al., 2012).

Data analysis

Data were analysed by applying the multilevel regression model (Bryk & Raudenbush, 1992; Hox, 2002). We started with the specification of the null, or no predictors model (Model 0) and continued by building two random intercept models with gender and the three dimensions of emotional intelligence at individual level (Model 1) and classroom level emotional intelligence (Model 2) as covariates. We then added two randomly varying slopes (Model 3) and finished by specifying a more complex random slopes and intercept model that tried to explain variability in the random slopes, including an interaction term between classroom level emotional intelligence and mood repair (Model 4).

For each model, the estimated values and standard errors of the fixed parameters and variance components were calculated, as was the deviance (the fit of the model). The difference between the deviance values was used to establish which of the models showed the best fit to the data. If the initial model constitutes a reduced version of the subsequent model, the difference between the deviance values follows a chi-squared distribution with as many degrees of freedom as the number of parameters in the extended model, minus the number in the reduced model.

Results

Table 1 presents the results of the multilevel regression models described above with depression score as the criterion variable. The first model, the intercept only model (Model 0), serves as a baseline; it showed that the total variance was

divided into two parts, 31.8 at student level and 4.61 at classroom level. Starting from this information the intra-class correlation coefficient, in other words, the proportion of variance accounted for at classroom level was calculated. This coefficient showed that approximately 13% of the total variability in depression scores occurred between classrooms. In the next model (Model 1), we can see that clarity of feelings and mood repair explained a substantial part of this variation at individual (23%) and at classroom (48%) level, substantially improving the fit of the model ($\Delta\chi^2/\Delta df = 163.64$; $p \leq 0.0001$). Both variables showed a negative relationship with depression (effect size values were $r = 0.18$ and $r = 0.41$ for clarity of feelings and mood repair, respectively). However, although the relationship between gender and depression ($r = 0.05$) and between attention to feelings and depression ($r = 0.02$) confirmed the expected tendency, the effect size was small in both cases. The fit also improved notably ($\Delta\chi^2/\Delta df = 13.84$; $p \leq 0.005$) when classroom level emotional intelligence was included in the model. This predictor variable had an important influence on depression score ($r = 0.34$) and reduced considerably the variance component at classroom level (18%). As expected, following the same pattern as in the case of individual level emotional intelligence, there was a negative relationship between the classroom's emotional intelligence and pupil's depression.

Since individual clarity of feelings and mood repair were significantly related to depression, in the next model we decided to test whether the slopes varied across classrooms. Model 3 showed that allowing slopes to vary randomly did not improve the fit of the model. However, the variance component for the mood repair-depression slope was statistically significant. The remainder of the fixed and random effects showed similar values as in the previous model. Finally, in model 4, we introduced a cross-level interaction between classroom level emotional intelligence and mood repair in order to examine whether the linear relationship between mood repair and depression changed according to the level of emotional intelligence in the classroom. The interaction term was not statistically significant in spite of producing a r value of 0.15. The variance components remained the same at individual and at group level. The fit of the model did not show any change. Thus, classroom level emotional intelligence slightly improved the effect of mood repair on depression but it did not show the significant effect we had anticipated.

Table 1. Results of the Multilevel Analyses for the sequence of models with gender and emotional intelligence at individual and group levels as predictor variables with depression as the criterion variable

Effect	Model 0	Model 1	Model 2	Model 3	Model 4
Fixed effects					
Intercept (γ00)	21.61 (0.23)	21.84 (0.21)	21.83 (0.20)	21.82 (0.20)	21.80 (0.20)
Pupil variables					
Male		-0.49*(0.22)	-0.48* (0.22)	-0.45*(0.22)	-0.44 (0.22)
Attention		0.01 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
Clarity		-0.18** (0.02)	-0.18** (0.02)	-0.19** (0.02)	-0.19** (0.02)
Repair		-0.44** (0.02)	-0.44** (0.02)	-0.44** (0.03)	-0.44** (0.02)
Classroom variables					
Group emotional intelligence (GEI)			-0.16** (0.04)	-0.15** (0.04)	-0.15** (0.04)
Interactions					
GEI x Repair					0.008 (0.006)
Variance components					
Within-subject (σ2 e)	31.80** (0.99)	24.57** (0.77)	24.57** (0.77)	23.83** (0.78)	23.86** (0.78)
Between-subjects (τ00) Intercepts	4.61** (0.84)	2.37** (0.50)	1.93** (0.45)	1.93** (0.45)	2.01** (0.46)
Between-subjects (τ21) Slopes				0.01 (0.01)	0.01 (0.01)
Between-subjects (τ31) Slopes				0.02* (0.01)	0.02* (0.01)
Model fit					
Model deviance					
Δ Deviance (1-0)	13847.82	13193.24	13179.40	13170.07	13167.64
Δ Deviance (2-1)		654.58**	13.84*	9.33	2.43
Δ Deviance (3-2)					
Δ Deviance (4-3)					
Δ df		4	1	2	1

Note. All predictor variables were centred over the grand mean. Standard errors are listed parenthetically; γ_{00} = Population mean of the average intercept; σ^2_e = Within-subject variance; τ_{00} = Variance of the intercepts (between-subjects variance); τ_{21} and τ_{31} = Variances of the slopes (between-subjects variances); *p<.05 **p<.0001

DISCUSSION

The results of this study have shown that high levels of emotional clarity and mood repair are related to lower levels of depression. These results are coherent with those reported in other studies (Fernández-Berrocal et al., 2006; Martinez-Pons,

1997) and extend previous results obtained with university students to adolescent population (Williams, Fernández-Berrocal, Extremera, Ramos, & Joiner, 2004).

As for the results of group EI and depression, we observed that there was a negative relationship between classroom EI and depressive symptoms. These results are coherent with those observed in the research of flow theory (Csikszentmihalyi & Csikszentmihalyi, 1988). However, we were unable to confirm that the relationship between individual EI and depression varied significantly according to group EI.

In respect to the relationships between gender and depression, the results showed that adolescent girls suffer from a slightly higher level of depression than boys.

Although the study is not based on ability measures and is of a correlational nature, we consider that it extends the incremental validity of self-perceived measures on the prediction of mental health and contributes to better explaining differences in adolescents' psychosocial adaptation when considering one of the most important contexts of reference: the classroom.

AUTHOR´S NOTE

The study was funded by a grant from the Science and Innovation Ministry of the Spanish Government (PSI2009-07280) and by a grant from the Research Bureau of the University of the Basque Country (General Funding for Research Groups GIU08/09 and Research Project EHU08/24).

REFERENCES

Abela, J. R. Z., Brozina, K., & Haigh, E. P. (2002). An examination of the response styles theory of depression in third- and seventh-grade children: A short-term longitudinal study. *Journal of Abnormal Child Psychology, 30,* 515-527.

Aritzeta, A., Balluerka, N., Alonso-Arbiol, I., Haranburu, M., Gorostiaga, A., & Gartzia, L. (in progress). *Group Emotional Intelligence: Development of a New Measure and its Associations with Gender and School Performance*.

Balluerka, N., Gorostiaga, A., & Haranburu, M. (2012). Validation of CDS (Children's Depression Scale) in the Basque-speaking population. *The Spanish Journal of Psychology, 15*.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models, Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

Csikszentmihalyi, M., & Csikszentmihalyi, I. (1988). *Optimal Experience: Psychological Studies of Flow in Consciousness*. New York: Cambridge University Press

Downey, L.A., Johnston, P.J., Hansen, K., Birney, J., & Stough, C. (2010). Investigating the mediating effects of emotional intelligence and coping on problem behaviours in adolescents. *Australian Journal of Psychology, 62*(1), 20-29.

Fernández-Berrocal, P., Salovey, P., Vera, A., Extremera, N., & Ramos, N. (2005). Cultural influences on the relation between perceived emotional intelligence and depression. *International Review of Social Psychology*, *18*, 91-107.

Fernández-Berrocal, P., Alcaide, R., Extremera, N. & Pizarro, D. (2006). The role of emotional intelligence in anxiety and depression among adolescents. *Individual Differences Research, 4*(1), 16-27.

Gorostiaga, A., Balluerka, N., Aritzeta, A., Haranburu, M., & Alonso-Arbiol, I. (2011). Measuring perceived emotional intelligence in adolescent population: Validation of the Short Trait Meta-Mood Scale (TMMS-23). *International Journal of Clinical and Health Psychology, 11*(3), 523-537.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion.* New York: Cambridge University Press.

Hox, J. J. (2002). *Multilevel Analysis. Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lang, M., & Tisher M. (1978). *Children's Depression Scale*. Camberwell, Victoria, Australia: Australian Council for Educational Research.

Martinez-Pons, M. (1997). The relation of emotional intelligence with selected areas of personal functioning. *Imagination, Cognition & Personality, 17*(1), 3-13.

Meyer, D. K., & Turner, J. C. (2006). Reconceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review, 18*, 377-390.

Nolen-Hoeksema, S., & Girgus, J.S. (1994). The emergence of gender differences in depression in adolescence. *Psychological Bulletin, 115*, 424-443.

Salguero, J. M., Fernández-Berrocal, P., Balluerka, N., Aritzeta, A. (2010). Measuring Perceived Emotional Intelligence in the Adolescent Population: Psychometric Properties of the Trait Meta-Mood Scale. *Social Behavior and Personality: An International Journal, 38*(9) 1197–1209.

Salovey, P. (2001). Applied emotional intelligence: Regulating emotions to become healthy, wealthy, and wise. In J. Ciarrochi, J. P. Forgas, & J. D. Mayer (Eds.), *Emotional Intelligence and Everyday Life* (pp. 168-184). New York: Psychology Press.

Skinner, E. A., & Zimmer-Gembeck, M. J. (2007). The development of doping. *Annual Review of Psychology*, *58*, 119-144.

Turner, J. C., Meyer, D. K., Cox, K. C., Logan, C., DiCintio, M., & Thomas, C. T. (1998). Creating contexts for involvement in mathematics. *Journal of Educational Psychology, 90,* 730-745.

Williams, F., Fernández-Berrocal, P., Extremera, N., Ramos, N., & Joiner, T. E. (2004). Mood regulation skill and the symptoms of endogenous and hopelessness depression. *Journal of Psychopathology and Behavioral Assessment, 26,* 233-240.

THE EFFECT OF THE ACTUAL SYSTEM OF PRACTICES ON EMPLOYEES' CITIZENSHIP BEHAVIOUR: A MULTILEVEL RANDOM COEFFICIENT ANALYSIS

Unai Elorza[1], Aitor Aritzeta[2] and Nekane Balluerka[2]
[1] Universidad de Mondragón
[2] Universidad del País Vasco
E-mail: uelorza@mondragon.edu

Abstract

More studies in the Strategic Human Resource Management (SHRM) field are required to explore how systems of practices that maximize a high commitment strategy are influencing employees' responses. This study aims to understand how the actual system implemented by managers has an effect on employees' citizenship behaviour. Multilevel results confirm that this relationship is mediated by the perceived system. Moreover, results show that the perceived system – citizenship slope is random among organizations and that this variability is partly explained by the actual system. These results point out the need to understand the shared meanings that are shaped collectively in different organizational contexts and how these meanings foster more favourable responses from employees.

There is a need in the SHRM field to understand how individual and group level phenomena are mediating between high commitment systems and organizational performance. More specifically, Wright and Boswell (2002) emphasized the need to research how systems of practices influence employees' responses. To date, there are few studies that combine in the same research both the actual system (from managers) and the perceived system (from employees) in order to understand their effect on individual and organizational level employee responses (Chang, 2005). The few studies that have analysed the effect of the system of practices on employees' responses, have followed a single informant approach at the individual level (Kehoe & Wright, 2010; Lam, Chen, & Takeuchi, 2009). So, it is necessary to measure the system from both perspectives (the actual and perceived one) in order to examine the effect of the system of practices on employees' attitudes and behaviour.

On the other hand, the SHRM field is focused almost exclusively on the content perspective of the system, not taking into account the social constructions that employees make about the system of practices (Bowen & Ostroff, 2004). The study developed by Nishii, Lepak and Schneider (2008) from the attribution point of view

is an exception in the field. It is necessary to explore to what extend the shared meanings that employees assign to the system of practices are influencing their response.

Thus, the aim of the study is twofold: (i) to understand how the actual system implemented by managers has an effect on the perceived system and on the citizenship behaviour of employees, and (ii) explore to what extend idiosyncratic interpretations of the perceived system in different organizations is leading to different citizenship behaviour responses.

HYPOTHESES DEVELOPMENT

Following Bowen and Ostroff's (2004) communication theory, the actual system of practices convey messages related to managers' commitment strategy. If the interpretation of the messages is consistent, employees' will likely exhibit the attitudes and behaviour desired by the organization. Thus, the effect of the actual system on employees' response might not be so straightforward because the ultimate behavioural response depends on the shared meaning that employees assign to the system rather than on the content of the system. The theoretical model that guide this study (see *Figure 1*) proposes that: (i) employees perceive the actual system (Liao, Toya, D P Lepak, & Hong, 2009), (ii) the perception of a high commitment system implies a favourable interpretation of managers' commitment leading to higher employees' commitment levels (Elorza, Aritzeta, & Ayestaran, 2011) and as a consequence to higher citizenship behaviour (Chang, 2005). The model also suggests that a perceived high commitment system does not necessarily mean a positive response on employees, because this response depends on the interpretation people make of this perceived system (Bowen & Ostroff, 2004). The shared meanings that are collectively construed depend not only on the perceived system but also on important variables related to the organizational context such as organizational culture.

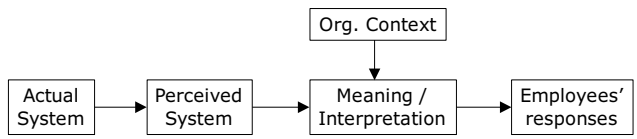


Figure 1. Theoretical model

Although the actual system does not necessarily mean that it is perceived by employees (Wright et al., 2001), there are studies that find a positive and significant relationship between both constructs (Elorza et al., 2011). Thus, the first hypothesis states that:

- Hypothesis 1: The greater the actual system, the higher the perceived system.

Higher levels of the perceived system are expected to increase employees' commitment (Kehoe & Wright, 2010), which will in turn have a positive effect on

citizenship behaviour (Podsakoff, Mckenzie, Paine, & Bachrach, 2000). Thus, it is expected a positive effect of the perceived system on the citizenship behaviour:

- Hypothesis 2: The greater the perceived system by employees, the greater their citizenship behaviour.

Bowen and Ostroff (2004) suggested that the interpretations employees make about the system might vary among different collectives and/or organizations due to distinctive collective sensemaking processes. That is, for the same content of the perceived system, different collectives (from the same and/or different organizations) might make diverse interpretations. As a result, employees shape different meanings about what the organization is expecting from them and therefore they exhibit diverse responses. The SHRM literature has assumed a constant and straightforward relationship between the system of practices and employees' responses. However, as a result of social interaction processes, employees' may give different meanings to the same perceived system in different organizational contexts. This phenomena should be visible through different relationships (slopes) between the system of practices and the behaviour of employees (different responses for the same perceived system). So, it is likely to find random slopes between different organizational contexts:

- Hypothesis 3: The effect of the perceived system on the citizenship behaviour is random among the different organizations.

The organizational context in which social interaction processes are held, have an important effect on the meaning that employees give to the perceived system. The actual system of practices is an important element of the organizational context from the employees' point of view. An organizational context where the actual system is high may be an indication of a «strong» commitment oriented context (Bowen & Ostroff, 2004). Thus, this study will use the actual system as a key element of the organizational context:

- Hypothesis 4: The perceived system – citizenship behaviour slope is partly explained by the actual system. That is, the higher the actual system the higher the slope.

METHOD

Sample

Twenty-six manufacturing plants of 25 different companies from the automotive, home appliance and machine tooling industries comprised the study sample. Of the 25 companies, 42% were public-limited companies, whereas the remainder were cooperatives co-owned by employees. Average size was 205 employees per unit. In order to improve the reliability of the study, the research focused only on shop-floor employees. The percentage of shop-floor employees per company that

took part in the study ranged from 20% to 100%, with an average figure of 54%, resulting in a total number of 1023 respondents.

Procedure

The final data were collected with a procedure consisting of two stages. First, two informants at managerial level (general manager and personnel director) were interviewed with a structured questionnaire in order to ensure the reliability of the actual system. And second, randomly selected managers' regular meetings with shop-floor employees were used to ask them about the perceived system and citizenship behaviour by means of another structured questionnaire. All the 1023 questionnaires gathered were examined in order to assess the quality of the answers. A final sample of 732 questionnaires from 26 manufacturing plants was included in subsequent statistical analyses.

Measures

Perceived system. The system is composed by training, participation, employment security, information sharing (financial and strategic), contingent compensation, autonomy and customer focus. A confirmatory factor analysis was performed to test a seven-factor model with 22 items. Fit indices showed an adequate overall fit: GFI = 0.93; AGFI = 0.91; TLI = 0.94; CFI = 0.95; RMSEA = 0.05. All item-factor loadings were statistically significant, and reliability indices confirmed the internal consistency of the seven constructs.

Citizenship behaviour. It was measured trough the scale proposed by Tsui, Pearce, Porter and Tripoli (1997). The confirmatory factor analysis yield overall good results: GFI = 0.98; AGFI = 0.96; TLI = 0.97; CFI = 0.98; RMSEA = 0.07. Composite reliability and extracted variance exceed the cut-off values.

Actual System. Items of the perceived system were adapted to be answered by managers in an interview format. Managers' responses were aggregated into a single value because the ICC(1,k) for the items and means exceeded the commonly accepted cut-off value of 0.70.

Control variables. Industry, size and ownership at the organizational level, and tenure and contract type at the individual level, were used as control variables in the study.

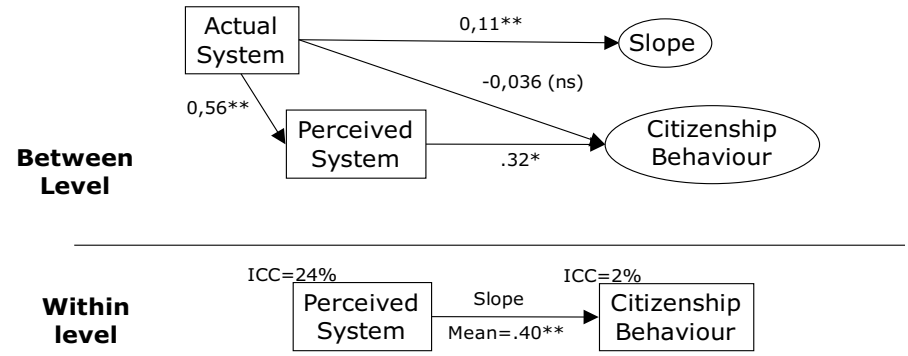
Data Analysis

Analyses were performed using Mplus with MLR estimator and residualized data (after controlling for the five control variables).Due to the multilevel nature of the data, multilevel SEM was used to test the hypotheses. Simple models are recommended when the group-level sample size is relatively small. Thus, average

values representing each construct rather than latent variables were used in the final model and observed variables were controlled in all five control variables following the residualized procedure outlined in Cohen and Cohen (1983).

Results

Figure 2 illustrates the final model tested. Unstardardized figures are displayed. Results show: (i) a positive and significant effect of the actual system on the perceived system (hypothesis 1), (ii) a significant relationship between the perceived system and citizenship behaviour at both within and between levels of analysis (hypothesis 2), (iii) a complete mediation of the perceived system at the between level, (iv) that the perceived system – citizenship behaviour slope is random among companies (hypothesis 3), and (v) that this variability is partly explained by the actual system (hypothesis 4).



* Satistically significant at p<.05
** Statistically significant at p<.01

Figure 2. Graphical representation of the tested model

Discussion

The research shows that the actual system has an influence on employees' citizenship behaviour through two paths. The first path is through its effect on the perceived system. The study demonstrates that the actual system has an effect on citizenship behaviour through the perceived system. Thus, the citizenship behaviour is closer to the perceived system rather than to the actual system. Almost all black box studies in the SHRM field measure either the actual system or the perceived one. However, results of the study show that when the aim is to understand the effect of the system on employees, the measurement of both constructs is a better approach.

The second path is through the perceived system – citizenship behaviour slope (see *Figure 3*). The results of the study show two evidences related to this second

path: (i) that the perceived system - citizenship slope is random among different companies and (ii) that the actual system partly explains its variability. The first evidence means that there can be different citizenship behaviour responses (in different organizational contexts) for the same content of the perceived system. The meaning that is given to the same perceived system value is different among organizations, and this meaning is influencing employees' citizenship behaviour. Almost all black box studies in the SHRM field are focused on a system content perspective. However, this study demonstrates that the same content can lead to different responses suggesting that the meaning that collectively is given to this content vary among organizations. Thus, the study is supporting Bowen and Ostroff's (2004) suggestion of incorporating in the SHRM research agenda the effect of the shared meanings (that are built up in collective sensemaking processes) on employees' responses.

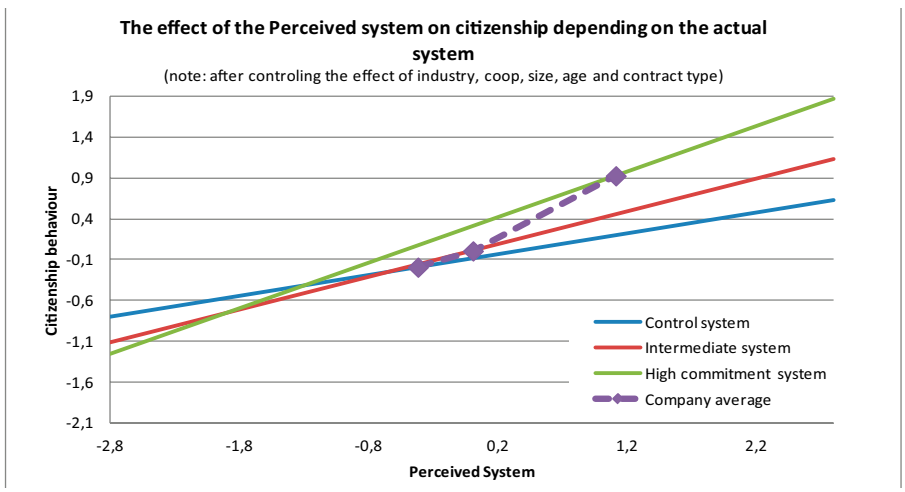


Figure 3. Different slopes depending on the actual system

Finally, the second evidence shows that the higher the actual system the higher the slope. This implies that the shared meaning that employees give to the perceived system is more favourable in high commitment contexts. However, this favourable meaning is related to the level of the perceived system. The highest and lowest citizenship behaviour can be obtained in high commitment contexts. If the perceived system is low in a high commitment context, the citizenship behaviour is the lowest and the other way round.

CONCLUSIONS

The study shows that the actual system elicits higher citizenship behaviours through the mediating role of the perceived system. However, it also shows that contrary to the main assumption in almost all studies in the SHRM field, the perceived system – citizenship slope is not constant among organizations. This implies

that the meaning that people give to the same perceived system content can vary among organizations influencing its effect on citizenship behaviour. Future black box studies should overcome the influence that attributions or shared meanings have on employees' responses.

REFERENCES

Bowen, D. E., & Ostroff, C. (2004). Understanding HRM-firm performance linkages: The role of the «strength» of the HRM system. *Academy of Management Review*, *29*(2), 203-221.

Chang, E. M. (2005). Employees' overall perception of HRM effectiveness. *Human Relations*, *58*(4), 523-544.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Elorza, U., Aritzeta, A., & Ayestaran, S. (2011). Exploring the black box in Spanish firms: the effect of the actual and perceived system on employees' commitment and organizational performance. *The International Journal of Human Resource Management*, *22*(7), 1401-1422.

Kehoe, R. R., & Wright, P. M. (2010). The Impact of High Performance Human Resource Practices on Employees' Attitudes and Behaviors. *Journal of Management*. doi:10.1177/0149206310365901

Lam, W., Chen, Z. G., & Takeuchi, N. (2009). Perceived human resource management practices and intention to leave of employees: the mediating role of organizational citizenship behaviour in a Sino-Japanese joint venture. *International Journal of Human Resource Management*, *20*(11), 2250-2270.

Liao, H., Toya, K., Lepak, D P, & Hong, Y. (2009). Do They See Eye to Eye? Management and Employee Perspectives of High-Performance Work Systems and Influence Processes on Service Quality. *Journal of Applied Psychology*, *94*(2), 371-391.

Nishii, L. H., Lepak, David P., & Schneider, B. (2008). Employee Attributions of the «Why» of HR Practices: Their Effects on Employee Attitudes and Behaviors, and Customer Satisfaction. *Personnel Psychology*, *61*(3), 503-545.

Podsakoff, P. M., Mckenzie, S., Paine, J., & Bachrach, D. (2000). Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management*, *26*(3), 513-563.

Tsui, A. S., Pearce, J. L., Porter, L. W., & Tripoli, A. M. (1997). Alternative approaches to the employee-organization relationship: Does investment in employees pay off? *Academy of Management Journal*, *40*(5), 1089-1121.

Wright, P. M., & Boswell, W. R. (2002). Desegregating HRM: A review and synthesis of micro and macro human resource management research. *Journal of Management*, *28*(3), 247-276.

Wright, P. M., Gardner, T. M., Moynihan, L. M., Park, H. J., Gerhart, B., & Delery, J. E. (2001). Measurement error in research on human resources and firm performance: Additional data and suggestions for future research. *Personnel Psychology*, *54*(4), 875-901.

A BIVARIATE LATENT CHANGE SCORE STRUCTURAL BUT DYNAMIC ANALYSIS OF LONGITUDINAL INTELLIGENCE DATA ON CHILDREN

John J. McArdle
University of Southern California
E-mail: jmcardle@usc.edu

Abstract

This is a brief methodological talk on *the use of simple linear dynamic expressions to deal with the possibility of complex changes*. This discussion will be limited to currently available analyses based on *Linear Structural Equation Models* (*LSEM;* see Jöreskog & Sörbom, 1979). Real data used here are based on summary scores from the *Weschsler Intelligence Scale for Children* (*WISC*; see McArdle & Epstein, 1987) where we pursue various kinds of multivariate analyses anyway to see what we can learn. Several questions are raised about the current practices in *latent growth models* or *multi-level modeling*. The latter methods opened up the possibilities of longitudinal modeling, but are seriously limited and do not capture all available changes. On the other hand, the *latent change score* (LCS) approach applied to common factors (as in McArdle & Nesselroade, 1994, 2003, 2012), does not advocate any particular position on the change model and this allows a wide variety of unique opportunities for model building, including an understanding of the sequential dynamics of relationships among different constructs. Graphic details of these methods will not be presented, but computer scripts (in R) and the WISC data are available so all analyses can be replicated.

There has been a dramatic rise in the use of latent variable modelling methods for longitudinal data analysis during the past decade (see McArdle, 1988, 2001, 2009; McArdle & Nesselroade, 2012). The most popular methods are known as *random coefficients*, or *multilevel models*, or *latent growth models*. While we see no general problem with their use, and we recognize that these are certainly improvements over the *Analysis of Variance of Repeated Measures* (see Bock, 1975), the limitations of these new methods are not often considered. Instead, these new methods are assumed to be useful for all new problems, and they certainly are not.

One basic problem arises when we have more than one outcome variable, and this is often the case in real longitudinal research. Only a few recent treatments (e.g., Horn, 1972; McArdle, 1988; Duncan, Duncan, & Strycker, 2006; Ferrer, Balluerka, & Widaman, 2008) have dealt with these key issues. But these presentations

seem to struggle to find the right model or are simply not very critical of the current approaches. Assuming we are limited by *linear structural equation models* (as in Jöreskog & Sorbom, 1979 we offer some *latent change score* (LCS) approach to open up the possibilities, and also to criticize other accounts.

The goal of this presentation is simply to apply alternative methods based on latent changes to the same data set to illustrate some possibilities. In succession we deal with models based on (1) curves of factor scores (CUFFS), (2) factors of curve scores (FOCUS), and (3) bivariate latent change scores models (BLCS). Most aspects of these models were previously presenetd and analyzed in papers on latent growth curve models (for details, see McArdle & Epstein, 1987; McArdle, 1988, 1989; McArdle & Aber, 1990; McArdle & Nesselroade, 1994, 2003, 2012). But the unique feature of this presentation is that we bring together all these different multivariate issues in the analysis of a single dataset.

METHOD

Participants

The longitudinal data used for the illustrations here are intellectual ability data from the *Wechsler Intelligence Scale for Children* (*WISC;* Wechsler, 1949) collected on a sample of N=204 children measured in a study by described by McArdle & Epstein (1987). At the initial testing, the children were in the first grade (at about age 6), and were measured on a wide variety of individual and group variables. The second testing (first retest) was about one year later when the children were in the second grade (about age 7), and many of the same measurements were repeated. The third testing was two years later, during the fourth grade (about age 9), and the fourth testing was another two years later, during the sixth grade (about age 11). One interesting aspect of these longitudinal data is that all records used here are fully complete. This was the way the data were given to the author by the data collector (in 1984). Of course, we recognize that this is an extremely rare event, but we will not need to deal with incompete data here (Hedeker & Gibbons, 1997; McArdle & Hamagami, 1991; McArdle et al., 2005).

Measurements

Multiple scores at each occasion are used to from a *Verbal* (V[t]) composite, formed as an unweighted average of each child's scores on four WISC «Verbal» subscales (Information, Comprehension, Similarities, and Vocabulary). Other scores measured at the same time are a Non-Verbal or Performance (P[t]) composite based on an unweighted average of four WISC «Performance» subscales (Picture Completion, Picture Arrangement, Block Design, and Object Assembly). For ease of presentation we have converted the original raw scores into a percent-correct metric from 0 to 100 (McArdle, 1988), but this rescaling is only used for con-

venience of interpretation and does not alter any of the psychometric or statistical features of the raw scores.

In the notation used here we assume the participants (n = 1 to 204) have been independently sampled on the same observed scores at some occasion-of measurement (m=1 to 4) coincident with some time-scale of interest (age or grade t=1 to 6). The scores are assumed to have been repeatedly measured under the same conditions and measured in the same units at all times. From such measurements we should be able to create a variety of statistical indicators describing the time-series, including observed means ($\boldsymbol{M}[t]$), standard deviations ($\boldsymbol{D}[t]$), and time-to-time correlations ($\boldsymbol{R}[t,t+j]$). This initial statistical summary of the WISC information on all N=204 children are presented in various other papers (and reproduced here in Appendix [1]). For example, the mean of Verbal composite at the first occasion is $\boldsymbol{M}[1]$=19.58 (i.e., 20% of the items answered correctly) while the mean at the second occasion is $\boldsymbol{M}[2]$=25.41 (i.e., 25% correct).

Models for Data Analysis

Data are analysed here using the laavan structural equation modelling program as coded in R (Rossel, 2006; see Appendix, and McArdle & Nesselroade, 2012, for computer scripts). Alternative multivariate models were presented based on three different starting points (defined below). For each model, the estimated values and standard errors of the fixed parameters and variance components were calculated, as was the deviance (i.e., L^2, the misfit of the model). If the initial model constitutes a reduced version of the subsequent model, and the residuals are normally distributed with zero means, then the difference between the deviance values can be summarized with a chi-squared distribution (χ^2) with as many degrees of freedom (df) as the number of parameters in the extended model, minus the number in the reduced model.

1. *Curve of Factor Scores* (CUFFS) models are described by McArdle (1988), Duncan, Duncan, and Strycker (2006) and Ferrer, Balluerka, and Widaman (2008). The basic idea is to fit a series of models of change in an invariant common factor score (f[t]). This reasonable sounding model unfortunately requires a rigid form of *metric factorial invariance* (MFI; following Meredith & Horn, 2001) among the measures. This means we need to establish a model where the factor loadings do not differ over time, so here we require

$$[1a] \qquad V[t]_n = \Lambda_v f[t]_n + uv_n \text{ and } P[t]_n = \Lambda_p f[t]_n + up_n ,$$

before we can examine any models of change in the common factors (f[t]). That is, if we cannot establish some form of MFI, we cannot go any further. As in any common factor model (even one with only two variables) we need to fix at least one of the factor loadings at a positive value to achieve identification. The unique means and variances of the variables can be considered as well, but it is the common factor and its changes which are of most interest. The model basically requires us to establish MFI (for a perspective on this topic, see McArdle & Cattell).

Once we do establish MFI we might consider applying a latent change model (after McArdle, 2001) to the common factor scores where we can combine elements of latent growth and autoregressive changes, such as

[1b] $$\Delta f[t]_n = (f[t]_n - f[t\text{-}1]_n) = \alpha_n + \beta\, f[t\text{-}1]_n + z_n\ .$$

This model allows the common factors to have a constant change (α), plus a proportional change, plus an independent innovation (z; typically represented with zero mean and constant variance ϕ^2_z). This model [1] can be fitted simultaneously or in parts.

2. *Factor of Curve Scores* (FOCUS) is an alternative multivariate model that was first used by McArdle (1988) but not used by many others since (see Hertzog et al, 2006). However, it is typical to think of this model now is a *parallel growth curve* framework (as in McArdle, 1989, following Meredith & Tisak, 1990) and that is exactly how it will be used here. In this model, each variable is presumed to reflect a latent growth process as

[2a] $$V[t] = g0_n + \Lambda[t]_v\, g1_n + uv_n \text{ and } P[t] = h0_n + \Lambda[t]_p\, h1_n + uv_n$$

where we have latent variables representing *initial levels* (*g0, h0*), and latent slopes (*g1*, *h1*) with potentially unknown factor loadings ($\Lambda[t]_v$, $\Lambda[t]_p$), and unique within time variances (*uv, up*). Although it is not often considered, there can also be a single covariance within time about the unique terms (ψ_{vp}).

The key hypotheses which can be examined here are important because they do not require metric factorial invariance. Instead, we can see if the change model for any variable follows (a) a prescribed pattern of change i.e., a linear Λ=[0,1, 3, 5]), or (b) the changes are latent (i.e., a latent Λ=[0,1, λ_3, λ_5]), or if the changes are parallel (i.e., $\Lambda[t]_v$=$\Lambda[t]_p$). But what is far more common is for researchers to hypothesize that the latent slopes are *correlated* (see Hertzog et al., 2006), and this then is simply a test of whether or not $\psi_{h1,g1}$ =0. But, instead, what we might really desire here is a test of whether the latent slopes are *connected.* This turns out to be related to multiple covariances of the different latent slopes, but this can be accomplished by requiring

[2b] $$\psi_{h1,g1} = \psi_{h0,g1} = \psi_{h1,g0} = 0.$$

Essentially, this three *df* constraint forces the slope of one to be unconnected to the slopes of the other. There are other ways to evaluate this hypothesis, but we start here.

3. *Bivariate Dual Change Score* (BDCS) model was first described by McArdle (2001) and has been used by many others (e.g., Ghisletta & Lindenberger, 2005; for review, see McArdle, 2009). Unlike the other models listed above, the BDCS model suggests that each variable represents it own unitary construct, but it has some measurement error, and these constructs may produce one another over time. To allow unique estimates of measurement error at least four occasions are required (Heise, 1973; Jöreskog & Sorbom, 1979), and we write

[3a] $$V[t]_n = g[t]_n + uv_n \text{ and } P[t]_n = h[t]_n + up_n ,$$

so we have a separation of the change components (in $g[t]$ and $h[t]$) from the unique components (uv and up). This also allows us to more easily define the latent changes as

[3b] $$\Delta v[t]_n = g[t]_n - g[t\text{-}1]_n \text{ and } \Delta p[t]_n = h[t]_n - h[t\text{-}1]_n$$

But when this is done we can then write a change model for each score by connecting the two series by introducing *cross-coupling coefficients* (γ) as

[3c] $$\Delta v[t]_n = \alpha_{vn} + \beta_v \, g[t\text{-}1]n + gv \; h[t\text{-}1]n + zvn \; and$$

$$Dp[t]n = apn + bp \; h[t\text{-}1]n + \gamma_p \, g[t\text{-}1]_n + zp_n .$$

It is somewhat surprising is that we can evaluate a model of this complexity with basic LSEM. But by using this approach the specific restrictions on the model parameters allows us to examine hypotheses about the leading and lagging latent indicators (for more details, see McArdle, 2001, 2009). This, of course, differs from a simple univariate form where some measured external variable (X) is used to account for variance in the latent slopes. This kind of model was used in a standard latent path regression model by McArdle & Epstein (1987), and also in many longitudinal applications of «multilevel» or «hierarchical» modeling (Bryk & Raudenbush, 1992; McArdle & Hamagami, 1996; Meredith & Horn, 2001). In these models any observed score X could be included directly or re-written as a latent true score. But here we assume X[t] is measured over time (see McArdle, 2007). This BLCS model was fitted to the WISC data by McArdle (2001) in an effort to evaluate the sequence of events in this series.

RESULTS

The LSEM factor-regression models described above will be fitted using the lavaan program in R to the two WISC composite scores (for details, see McArdle & Nesselroade, 2012). The unequal spacing between these occasions was handled in the same way with all models (i.e., with the two incomplete grades $Y[3]$ and $Y[5]$) and all models were fitted to the means, deviations, and correlations of the raw scores. A model of unrestricted means and covariances fitted to the four occasion WISC data yields a L^2=-5110 and this will be our basic comparison model. This basic fit is altered by a χ^2=1424 with df=28 if we say that the covariances are zero. An initial no-change baseline model used only 7 parameters to define a model with equal means, equal variances, equal within variable correlations, and an equal within variable correlation at all occasions. This restrictive model led to a very poor fit to the 44 summary statistics ($\chi^2\{37\}$=2807). Obviously, this model of no-change is not a very interesting model, but it is a model we must improve upon. Various highlights of the three alternative multivariate model discussed earlier will be presented.

CUFFS Results

The first set of models fitted used structural Equations [1] as a guide yields invariant factor loadings (Λ=1, 1.15), with α=7.2, and β=-0.13, and ψ_z^2=6.4. This could represent an interesting latent change model. Unfortunately, this model yields an overall fit of L^2=-5247 to the four occasion WISC data. When this is translated to a χ^2=276 with *df*=24 and ε_a=.227 (or just $\chi^2\{24\}$=276), we conclude this model of invariant factors does not fit very well. By relaxing the factor loading constraint we can improve the model a great deal (by $d\chi^2$=139 with *df*=3), so this is probably where our model fitting problem can be found. Now to be clear, we have already allowed free unique means and variances over time, so basic scaling is probably not the issue. Of course, many more models could be fit to try to understand this problem. But there actually seems like very little we can do at this point to evaluate the latent changes in the factors, because these result suggests the common factor is not scaled the same way at each occasion. And, in fact, this does not seem to be an atypical result. But perhaps we should not be surprised that the Verbal and Performance measures do not work together in the same way over all occasions. If so, this would be evidence for a General factor of intelligence, and we have never found such evidence (see McArdle, et al., 2002; McArdle, 2007).

FOCUS Results

The second set of models fit used structural Equations [2] as a guide yields and yielded changing factor loadings (Λ_v=[=0, =1, 2.3, 4.3], and Λ_p=[=0, =1, 2.2, 3.4]), with μ_{gl}=5.6 and μ_{hl}=9.8, and with ϕ_{gl}^2=2.5 and ϕ_{hl}^2=3.3, and the covariance $\phi_{gl, hl}$=1.1. This model yields an overall fit of L^2=-5132, and this yields $\chi^2\{23\}$=45 with ε_a=.069. So we conclude this model of nonlinear shaped latent growth factors does in fact fit very well. By requiring the within unique covariance to be zero we do not reduce the model fit a great deal (by $d\chi^2$=5 with *df*=1), so this is probably not where any fitting issues can be found. But by requiring the factor loading constraints to both be linear we reduce the model fit a great deal (by $d\chi^2$=56 with *df*=4), so this is probably not where any fitting issues can be found. Furthermore, by requiring the factor covariance constraints to be zero (as in [2b]) we also reduce the model fit a great deal (by $d\chi^2$=64 with *df*=3), so we conclude the nonlinear shaped factors are connected. Obviously there are some specific changes over time within each variable, and these changes may even be connected, and this model does not require any form of MFI. Unfortunately, it is very hard to tell anything about the sequence of events using this FOCUS approach.

BLCS Results

The next model fit used structural Equations [3] as a guide. Although there is a lot to say about this model, only the bare minimum will be presented here. In previous work (McArdle, 2001) each variable was initially dealt with separately, and

this univariate approach still seems like a good strategy to follow. But for now, we will start with the bivariate result and move onto the specific tests of hypotheses. To simplify these models the equations of change are presumed to be invariant over time. This form of invariance offers many interesting and testable hypotheses, especially with more occasions of measurement. But here, to simplify these issues here, we report on only three models: (a) the overall model, (b) a model where there is no V → P, and (c) a model where there is no P → V.

Our first model is a fully saturated BDCS model. As starting values for these SEM we used the MLE obtained in the previously separate univariate DCS models. In contrast to the univariate results for Verbal (V[t]) scores, here we first obtain level variance (ϕ_{g0}^2=25) and unique variance (ψ_v^2=11), and this implies a reliability (at time 1) of about (η_v^2=0.69). In contrast, the model for Performance (*P*[t]) here has level variance (ϕ_{h0}^2=57) and unique variance (ψ_p^2=22), and this implies a reliability (at time 1) of about (η_p^2=0.72). This fully-saturated model includes 21 parameters and the fit (χ^2{23}=75, e_a=.11) is a relatively decent fit and a large improvement over the baseline no-change model. The resulting *dynamic but structural equations* for a fully saturated BDCS model can be written as

$$[4a] \quad E\{\ddot{A}v[t]_n\} = \alpha_{g0}\,\mu_{g0} + \beta_v\,E\{v[t\text{-}1]_n\} + \gamma_v\,E\{p[t\text{-}1]_n\}$$
$$= -2.98 + 0.77\,E\{v[t\text{-}1]_n\} + -0.44\,E\{p[t\text{-}1]_n\}$$

$$[4b] \quad E\{\ddot{A}p[t]_n\} = \alpha_{h0}\,\mu_{h0} + \beta_p\,E\{p[t\text{-}1]_n\} + \gamma_p\,E\{v[t\text{-}1]_n\}$$
$$= 8.1 + -0.38\,E\{v[t\text{-}1]_n\} + 0.38\,E\{p[t\text{-}1]_n\}$$

and it is clear that no innovation variance was considered here (but see McArdle, 2001). Now to test the alternative hypothesis that there is no P → V we constrain the cross-coupling of P on V (γ_v=0) and this yields a loss of fit of $d\chi^2$=25 with *df*=1. This relatively large loss of fit means we need to retain this parameter. The alternative of no V → P (γ_p=0) yields a loss of fit of $d\chi^2$=4 with *df*=1). This relatively small loss of fit means we can drop this parameter. We note that the removal of all cross-couplings (γ_v=γ_p=0) yields a loss of fit of $d\chi^2$=26 with *df*=2. Under these constraints, P → V is the important sequence parameter.

DISCUSSION

The results of this brief review are intended to illustrate how that many forms of longitudinal change models are possible with multivariate data. This is true even after the analyst will have chosen a level of measurement appropriate for their selected problem. Among several problems with the use of SEM advocated here are: (1) it only works with *apriori* hypotheses about the changes, and these are not often readily available, and (2) LSEM tests of goodness-of-fit assume normally distributed residuals. There are solutions for both of these persistent problems but none of these were isolated here (but see Wang & McArdle, 2008; McArdle, 2012).

From the perspective of the WISC intellectual ability constructs, it appears that we get different results from different models of the same longitudinal data. This

should be no surprise, but it often is (see Ferrer & McArdle, 2003). From the CUFFS approach we find that change is not possible to analyze because the common factor of these two variables changes its interpretation or meaning at each time. From the FOCUS approach we can see that each variable probably has its own latent change model, and these are not linear shapes and they are connected. From the BDCS approach the Non-Verbal (or Performance) components are the leading indicators of the Verbal components, and it is not the other way around. It seems to us that the BCDS approach offers the most appealing results.

Actually, the P→V relationship is complicated by the negative impact – thus, if anything, it appears that the Performance factor has a limiting role in the growth of the Verbal factor. This does not really seem like the «Investment Theory» predictions (Ferrer & McArdle, 2004), and this result seems like more of the opposite impact. Of course, we also realize these WISC data are observational data where little is under experimental control, and this longitudinal result only gives us a starting point (e.g., see Shadish, Cook, & Campbell, 2002; McArdle, 2006; Shrout, 2011). This result now needs be examined further by experimentation where the Performance factor is under experimental contrail, and P is raised and P is lowered. If this manipulation cannot be done, for practical or ethical reasons, then we are only left with these kinds of structural inferences about the dynamic sequence.

There are certainly other multivariate approaches which can be taken (see McArdle, Hamagami, Bradway, & Meredith, 2000; McArdle & Nesselroade, 2012) but the three approaches used here illustrate a variety of contemporary solutions.

AUTHOR'S NOTE

The work described here has been supported since 1980 by the National Institute on Aging (Grant#AG-07137). I am especially grateful to the work of my close friend and colleague, Fumiaki Hamagami, and to my close collaboration with John Nesselroade. This research was also helped by the support of my many friends and colleagues, including Steven Aggen, Steven Boker, Emilio Ferrer-Caja, Paolo Ghisletta, John Horn, Bill Meredith, Carol Prescott, Keith Widaman and Dick Woodcock. Reprints can be obtained from the author at the Horn Psychometric Laboratory, Room 704 SGM Building, Department of Psychology, University of Southern California, Los Angeles, CA 90089 USA. Computer program input scripts used here can be found on our website → http://kiptron.psych.usc.edu.

REFERENCES

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Duncan, T.E., Duncan, S.C., & Strycker, L.A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd Ed). Mahwah, NJ: Erlbaum.

Ferrer, E., Hamagami, F., & McArdle, J.J. (2004). Modeling latent growth curves with incomplete data using different types of structural equation modeling and multilevel software. *Structural Equation Modeling*, *11* (3), 452-483.

Ferrer, E., Balluerka, N., & Widaman, K.F. (2008). Factorial invariance at the specification of second-order growth models. *Measurement, 4* (1), 22-36.

Ferrer, E., & McArdle, J.J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling*. *10*, 493-524.

Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology, 40*, 935-952.

Ferrer, E., McArdle, J. J., Shaywitz, B. A., Holahan, J. M., Marchione, K., & Shaywitz, S. E. (2007). Longitudinal models of developmental dynamics between reading and cognition from childhood to adolescence. *Developmental Psychology, 43*, 1460-1473.

Ghisletta, P. & Lindenberger, U. (2005). Exploring the structural dynamics of the link between sensory and cognitive functioning in old age: Longitudinal evidence from the Berlin Aging Study. *Intelligence, 33*, 555-587.

Ghisletta, P., & McArdle, J.J. (2012 expected). Teacher's Corner: Latent Change Score Models Estimated in R. *Structural Equation Modeling*.

Hamagami, F., McArdle, J.J., & Cohen, P. (2000). Bivariate dynamic systems analyses based on a latent difference score approach for personality disorder ratings. In V.J. Molfese & D.L. Molfese (Eds.). *Temperament and personality development across the life span*. Mahwah, NJ: Erlbaum.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64-78.

Heise, D.R. (1973). *Causal Analysis*. NY: Wiley

Hertzog, C., Lindenberger, U., Ghisletta, P., & von Oertzen, T. (2006). On the power of multivariate latent growth curve models to detect correlated change. *Psychological Methods, 11*(3), 244-252.

Horn, J.L. (1972). State, trait, and change dimensions of intelligence. *The British Journal of Mathematical and Statistical Psychology, 42* (2), 159-185.

Jöreskog, K.G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books

McArdle, J.J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics, 16* (1), 163-200.

McArdle, J.J. (1988). Dynamic but structural equation modeling of repeated measures data. In J.R. Nesselroade & R.B. Cattell (Eds.), *The Handbook of Multivariate Experimental Psychology, Volume 2*. New York, Plenum Press, 561-614.

McArdle, J.J. (1989). Structural modeling experiments using multiple growth functions. In P. Ackerman, R. Kanfer & R. Cudeck (Eds.), *Learning and Individual Differences: Abilities, Motivation, and Methodology*. Hillsdale, NJ: Erlbaum, 71-117.

McArdle, J.J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.). *Structural Equation Modeling: Present and future.* Lincolnwood, IL: SSI. 342-380.

McArdle, J.J. (2006). Dynamic Structural Equation Modeling in Longitudinal Experimental Studies. In Kees van Montfort, H. Oud & A. Satorra (Editors), *Longitudinal models in the behavioural and related sciences* (pp. 159-187). EAM Book Series. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McArdle, J.J. (2007). Five Steps in the Structural Factor Analysis of Longitudinal Data. In R. Cudeck & R. MacCallum, (Eds.). *Factor Analysis at 100 years* (pp.99-130). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McArdle, J.J. (2009). Latent variable modeling of longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McArdle, J.J. (2012). Exploratory data mining using CART in the Behavioral science. In H. Cooper & A. Panter (Eds.) *Handbook of Methodology in the Behavioral Sciences* (Chapter 20). American Psychological Association, Washington, DC: APA Books.

McArdle, J.J., & Cattell, R.B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research, 29* (1), 63-113.

McArdle, J.J., & Epstein, D.B. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58* (1), 110-133.

McArdle, J.J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R.W. (2002). Comparative longitudinal multilevel structural analyses of the growth and decline of multiple intellectual abilities over the life-span. *Developmental Psychology, 38* (1) 115-142.

McArdle, J.J. & Hamagami, F. (1991). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. In L. Collins & J.L. Horn (Eds.), *Best Methods for the Analysis of Change*. Washington, D.C.: APA Press, 276-304.

McArdle, J.J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcoulides & R. Schumacker (Eds.), *Advanced Structural Equation Modeling Techniques*. Hillsdale, N.J.: Erlbaum. 89-124.

McArdle, J.J., & Hamagami, F. (2004). Longitudinal tests of dynamic hypotheses on intellectual abilities measured over sixty years. In S.M. Boker, et al. (Editors), *Quantitative Methods in Contemporary Psychology.* Mahwah: Erlbaum.

McArdle, J.J., Hamagami, F., Meredith, W., & Bradway, K.P. (2000). Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learning and Individual Differences*, *12* (2000), 53-79.

McArdle, J.J., & Nesselroade, J.R. (1994). Using multivariate data to structure developmental change. In S.H. Cohen & H.W. Reese (Eds.), *Life-Span Developmental Psychology: Methodological Innovations*. Hillsdale, N.J.: Erlbaum, 223-267.

McArdle, J.J., & Nesselroade, J.R. (2003). Growth curve analyses in contemporary psychological research. In J. Schinka, & W. Velicer (Editors), *Comprehensive Handbook of Psychology, Volume Two* (pp. 447-480). NY: Pergamon Press.

McArdle, J.J., & Nesselroade, J.R. (2012). *Longitudinal Structural Equation Modeling.* Washington, DC: American Psychological Association Books.

McArdle, J. J., Small, B.J., Backman, L., & Fratiglioni, L. (2005). Longitudinal models of growth and survival applied to the early detection of Alzheimer's Disease. *Journal of Geriatric Psychiatry and Neurology, 18* (4), 234-241.

Meredith, W., & Horn, J.L. (2001). The role of factorial invariance in modeling growth and change. In L.M. Collins, & A. Sayers (Eds). *New methods for the analysis of change.* American Psychological Association, Washington DC: APA Books.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Rossel, Y. (2006). *The lavaan computer program for structural equations*. Unpublished Manuscript, University of Ghent: Ghent, Holland.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental. & Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Shrout, P. (2010). Integrating causal analysis into psychopathology research. (pp. 3-24). In Shrout, P.E. Keyes, K.M. & Ornstein, K. (eds.). *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures.* NY: Oxford University Press.

Wang, L., & McArdle, J. J. (2008). A simulation study comparison of Bayesian estimation with conventional methods for estimating unknown change points. *Structural Equation Modeling, 1*5(1), 52-74.

Wechsler, D. (1949). *Wechsler intelligence scale for children*. New York: The Psychological Corporation.

AVANCES METODOLÓGICOS EN META-ANÁLISIS

Coordinador: Julio Sánchez Meca
Universidad de Murcia

El meta-análisis es una metodología de investigación por la que un conjunto de estudios empíricos sobre un mismo tema se integran estadísticamente con objeto de obtener una estimación global del tamaño del efecto, comprobar si los efectos individuales son homogéneos en torno a su media y, en caso contrario, examinar el influjo de variables moderadoras de los estudios sobre dichos efectos. El propósito de este simposio es presentar las últimas investigaciones que venimos realizando en la Unidad de Meta-análisis de la Universidad de Murcia sobre algunos aspectos metodológicos del meta-análisis. El simposio está formado por cuatro comunicaciones, de forma que las dos primeras se centran en avances estadísticos mientras que las dos últimas se centran en un aspecto conceptual muy importante en meta-análisis, como es el de la valoración de la calidad metodológica de los estudios primarios. Las dos primeras comunicaciones presentan los resultados de sendos estudios de simulación Monte Carlo en los que se analizan las propiedades estadísticas de diversos procedimientos de análisis en el contexto meta-analítico. Así, en la primera comunicación José A. López López presenta un estudio comparativo mediante simulación Monte Carlo de varios métodos propuestos en la literatura para promediar un conjunto de tamaños del efecto, asumiendo diferentes modelos estadísticos, tales como el modelo de efectos fijos, el de efectos aleatorios, el modelo de coeficientes variables recientemente propuesto por Bonett y el modelo de efectos aleatorios propuesto por Hunter y Schmidt (2004). En la segunda comunicación Fulgencio Marín Martínez presenta los resultados de otro estudio de simulación en el que compara la adecuación de diferentes métodos para estimar la proporción de varianza explicada cuando se aplican modelos de meta-regresión de efectos mixtos en el contexto de un meta-análisis, para analizar el influjo de variables moderadoras de los tamaños del efecto. Las otras dos comunicaciones tienen que ver con el problema de cómo evaluar la calidad metodológica de los estudios primarios y cómo ésta puede estar afectando a las estimaciones de los efectos en la población. En esta línea, la tercera comunicación será presentada por José A. López Pina y en ella se presenta un estudio empírico de la validez y la fiabilidad de una escala de calidad metodológica elaborada por nuestro equipo de investigación y que todavía no había sido sometida a un análisis psicométrico profundo. Dicha escala está compuesta por diez ítems dicotómicos que se puntúan 1/0, dando lugar a una puntuación global de calidad mediante la suma de dichos ítems. El estudio psicométrico se ha realizado sobre los resultados de un meta-análisis realizado por

nuestro equipo de investigación. En la cuarta y última comunicación Julio Sánchez Meca presentará un estudio empírico en el que se comprueba cómo la calidad metodológica de los estudios primarios puede estar afectando a los tamaños del efecto obtenidos en los estudios y, en consecuencia, provocando sesgos en las estimaciones de los efectos en la población. Para ello, la misma escala de calidad compuesta por diez ítems dicotómicos mencionada antes fue puesta en relación con los tamaños del efecto de un meta-análisis realizado por nuestro equipo de investigación, para comprobar la existencia de tales sesgos en las estimaciones de los efectos.

PALABRAS CLAVE: Meta-Análisis, Modelos estadísticos, Meta-regresión, Calidad metodológica.

INTERVALOS DE CONFIANZA PARA EL EFECTO MEDIO EN META-ANÁLISIS: UNA COMPARACIÓN DE LAS ALTERNATIVAS EXISTENTES MEDIANTE SIMULACIÓN MONTE CARLO

José Antonio López-López[1], Wim van den Noortgate[2] y Fulgencio Marín-Martínez[1]

[1] Universidad de Murcia

[2] Universidad Católica de Lovaina

Correo electrónico: josealopezlopez@um.es

Resumen

Uno de los principales objetivos en meta-análisis consiste en la estimación del efecto medio, junto con su intervalo de confianza. Para ello, se pueden emplear distintos métodos, derivados de diferentes modelos estadísticos que plantean supuestos específicos y condicionan la generalizabilidad de los resultados. En este estudio se comparó el funcionamiento de seis métodos, en términos de cobertura del intervalo de confianza para el efecto medio, mediante simulación Monte Carlo. Entre las condiciones manipuladas, se incluyeron dos factores que amenazan la validez del modelo de efectos aleatorios, que es el más frecuentemente asumido en la actualidad. Estos factores son el cumplimiento de normalidad en la distribución de efectos paramétricos y la existencia de una correlación entre tamaño muestral y efecto. Los resultados mostraron un deterioro generalizado del funcionamiento de los métodos cuando alguno de estos factores estaba presente. El modelo de efectos aleatorios corregido por Hartung (1999) mostró un funcionamiento apropiado en un mayor número de condiciones que las demás alternativas.

El meta-análisis constituye en la actualidad una herramienta de gran utilidad para la eficiente acumulación del conocimiento científico y su aprovechamiento en el contexto aplicado (Sánchez-Meca, Marín-Martínez y López-López, 2011). En un meta-análisis, los resultados de cada estudio se convierten a una métrica común a través de un índice del tamaño del efecto. Uno de los objetivos fundamentales en un meta-análisis consiste en calcular una estimación puntual del tamaño del efecto medio, así como su intervalo de confianza.

Existen dos modelos estadísticos fundamentales en meta-análisis para acometer este objetivo. El primero de ellos es el modelo de efecto fijo, que asume un único efecto paramétrico para todos los estudios y cuyos resultados son sólo generalizables a los estudios incluidos en el meta-análisis o a otros muy similares. Por

otra parte, el modelo de efectos aleatorios (EA) asume una población de efectos paramétricos, de los cuales se selecciona aleatoriamente una muestra de efectos paramétricos que serán estimados por los respectivos estudios empíricos del meta-análisis.

El modelo EA es el más empleado por los meta-analistas en la actualidad, principalmente porque es el único que permite generalizar los resultados y conclusiones más allá de la muestra de estudios incluidos en el meta-análisis (Borenstein, Hedges, Higgins y Rothstein, 2010). Igual que en la mayoría de las investigaciones científicas, el objetivo en meta-análisis suele ser el conocimiento acumulativo, es decir, conocimiento generalizable (Schmidt, Oh y Hayes, 2009), por lo que existe un consenso general de que el modelo EA es el más apropiado en la mayoría de situaciones. Sin embargo, el modelo EA también tiene sus detractores. Bonett (2009) subrayó que en la práctica no existe muestreo aleatorio de los estudios, por lo que asumir una distribución subyacente concreta para los efectos paramétricos es difícil de justificar. Además, la estimación de la varianza inter-estudios, τ^2, puede no ser precisa cuando se dispone de pocos estudios, algo que, según Bonett, es habitual en meta-análisis. Finalmente, Shuster (2010) ha argumentado recientemente que es habitual que exista una correlación entre tamaño muestral y efecto. Dado que el factor de ponderación en EA (y en efecto fijo) es una función del tamaño muestral, esto podría suponer un sesgo en los resultados.

Basándose en las limitaciones técnicas y conceptuales del modelo EA, Bonett (2009) ha recomendado el uso de una opción intermedia, el modelo de coeficientes variables, previamente propuesto en meta-análisis (Laird y Mosteller, 1990), que comparte elementos de los dos anteriores, ya que asume que cada estudio estima un efecto paramétrico diferente, pero se considera que los efectos paramétricos no han sido extraídos al azar y, en consecuencia, sólo es apropiada la generalización de los resultados al grupo de estudios seleccionados o a otros idénticos.

En este estudio se compararon mediante simulación Monte Carlo seis procedimientos para calcular un intervalo de confianza en torno al efecto medio en meta-análisis, empleando diferencias medias tipificadas como índice del tamaño del efecto. En primer lugar, se incluyó el modelo de efecto fijo, ponderando por la inversa de la varianza muestral de cada estudio (Hedges y Vevea, 1998). También se incluyó el modelo de coeficientes variables, siguiendo la propuesta de Bonett (2009) para diferencias medias tipificadas, donde se asume que todos los efectos paramétricos son igualmente representativos y por ello no se ponderan. Además, varios procedimientos basados en un modelo EA, en los que se añade una segunda fuente de variación, τ^2, fueron considerados en este estudio. En concreto, se incluyó el método tradicional en un modelo EA (Hedges y Vevea, 1998), la corrección propuesta por Hartung (1999), la reciente aportación de Henmi y Copas (2010) para contrarrestar los efectos de una correlación entre tamaño muestral y efecto y, por último, el método de ponderación por el tamaño muestral (n) derivado de los trabajos de Hunter y Schmidt (2004).

En cuanto a nuestras hipótesis, el trabajo de Sánchez-Meca y Marín-Martínez (2008), en el que se compararon algunos de los métodos presentados aquí, nos llevaba a esperar que el método basado en un modelo de efecto fijo sólo funcionaría adecuadamente en las condiciones de efecto paramétrico común, mientras que el método EA que mostraría un mejor funcionamiento sería el basado en la corrección de Hartung (1999). No obstante, era esperado que todos los métodos EA, que requieren el cumplimiento del supuesto de normalidad para la distribución de efectos paramétricos, se verían afectados negativamente cuando esto no se cumpliera. Además, teniendo en cuenta que el valor de referencia en nuestra simulación era el hiperparámetro, o media de la distribución de efectos paramétricos, esperábamos un funcionamiento pobre para el método de coeficientes variables, que no asume muestreo aleatorio de efectos y, por tanto, tiene como hiperparámetro de referencia la simple media aritmética de los efectos paramétricos estimados en el meta-análisis.

MÉTODO

Se llevó a cabo un estudio de simulación, manipulando varias condiciones para crear una amplia variedad de escenarios realistas en meta-análisis. El número de estudios para cada meta-análisis fue una de las condiciones manipuladas, con valores $k = \{5, 10, 20, 40, 60\}$. Los k efectos paramétricos fueron generados con un promedio de 0,5. τ^2 también se manipuló con valores 0, 0,08 y 0,24, de manera que el primer valor implicaba homogeneidad en los efectos paramétricos, mientras que para los dos restantes los efectos fueron generados a partir de una distribución normal o log-normal.

Otra condición manipulada fue el tamaño muestral de cada estudio, generado a partir de una distribucíon chi-cuadrado, con valores promedio de 15, 25 y 50 sujetos. Finalmente, se manipuló la correlación entre tamaño muestral y tamaño del efecto, con valores 0, -0,15, -0,25, y -0,45. La comparación de los distintos métodos estadísticos se llevó a cabo a través del cálculo del ajuste empírico al nivel de confianza (95%).

RESULTADOS

Los datos de esta sección se resumen en cinco figuras, cada una de ellas formada por doce gráficos. Cada figura presenta los resultados dada una distribución de los efectos paramétricos (normal o log-normal) y un valor para τ^2, Además, cada gráfico refleja la cobertura del intervalo de confianza en el eje de ordenadas, mostrando el cambio producido por el número de estudios (abscisas), mientras que otros dos parámetros, el tamaño muestral promedio (N) y la correlación entre tamaño muestral y efecto (Cor) se mantuvieron fijos dentro de cada gráfico. Las líneas y puntos del interior de los gráficos reflejan el funcionamiento de los métodos, representados según recoge la Tabla 1.

Tabla 1. Leyenda común para todos los gráficos de este trabajo

Efecto fijo		EA: ponderación *n*	
EA tradicional		Coeficientes variables	
EA: corrección de Hartung		EA: propuesta de Henmi-Copas	

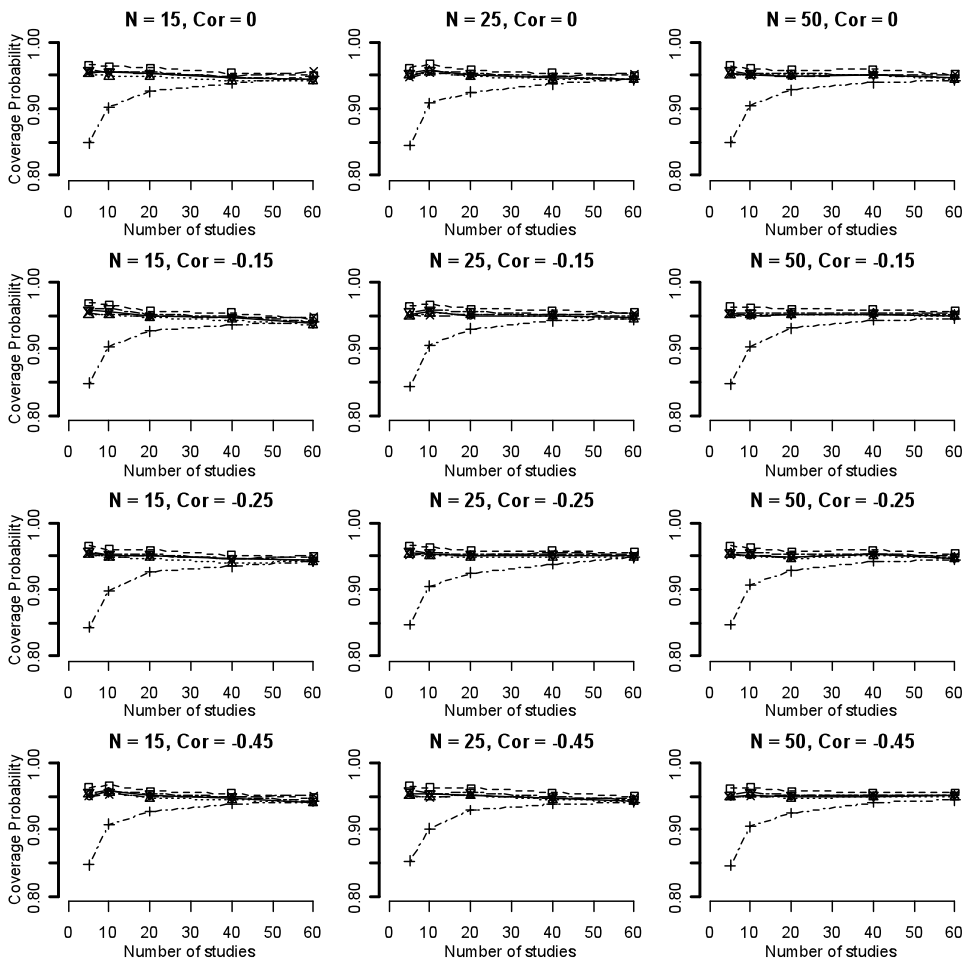


Figura 1. Cobertura de intervalo para las condiciones con $\tau^2 = 0$

La Figura 1 muestra los resultados cuando no existía distribución subyacente de los efectos paramétricos, es decir, cuando todos los estudios estimaban un mismo efecto, cumpliéndose así el supuesto del modelo de efecto fijo. Teniendo en cuenta que valores próximos a 0,95 indican un buen funcionamiento del método en

términos de cobertura del intervalo, sólo el método EA ponderando por *n* mostró un funcionamiento inapropiado, mejorando a medida que aumentaba el número de estudios.

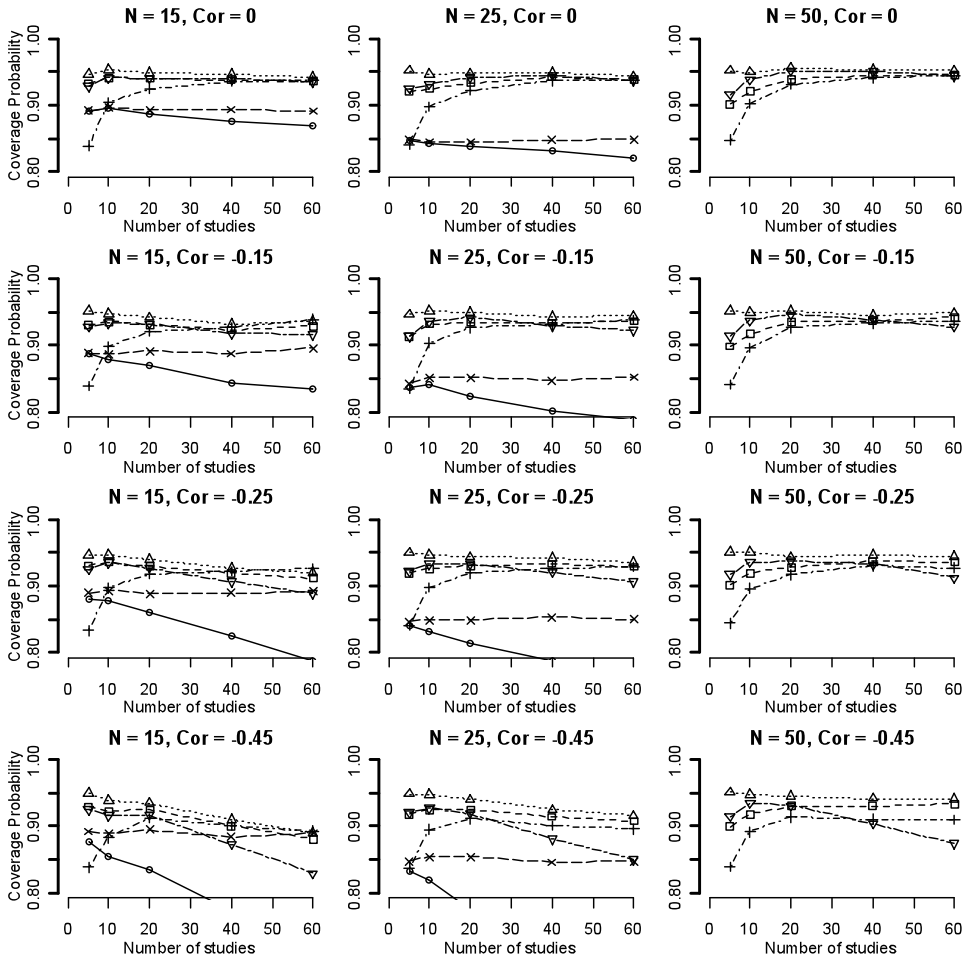


Figura 2. Cobertura de intervalo para distribución normal y $\tau^2 = 0{,}08$

En las condiciones con heterogeneidad entre los tamaños del efecto paramétricos (Figuras 2-5), los valores de cobertura para los modelos de efecto fijo y de coeficientes variables disminuyeron pronunciadamente, llegando en muchos casos a alcanzar coberturas por debajo de 0,80 (y por tanto, fuera del área representada en cada gráfico). En general, las tasas de cobertura disminuyeron para todos los métodos a medida que aumentó la correlación entre tamaño muestral y tamaño del efecto (Figura 2), siendo este hecho menos pronunciado para EA aplicando la corrección de Hartung (1999).

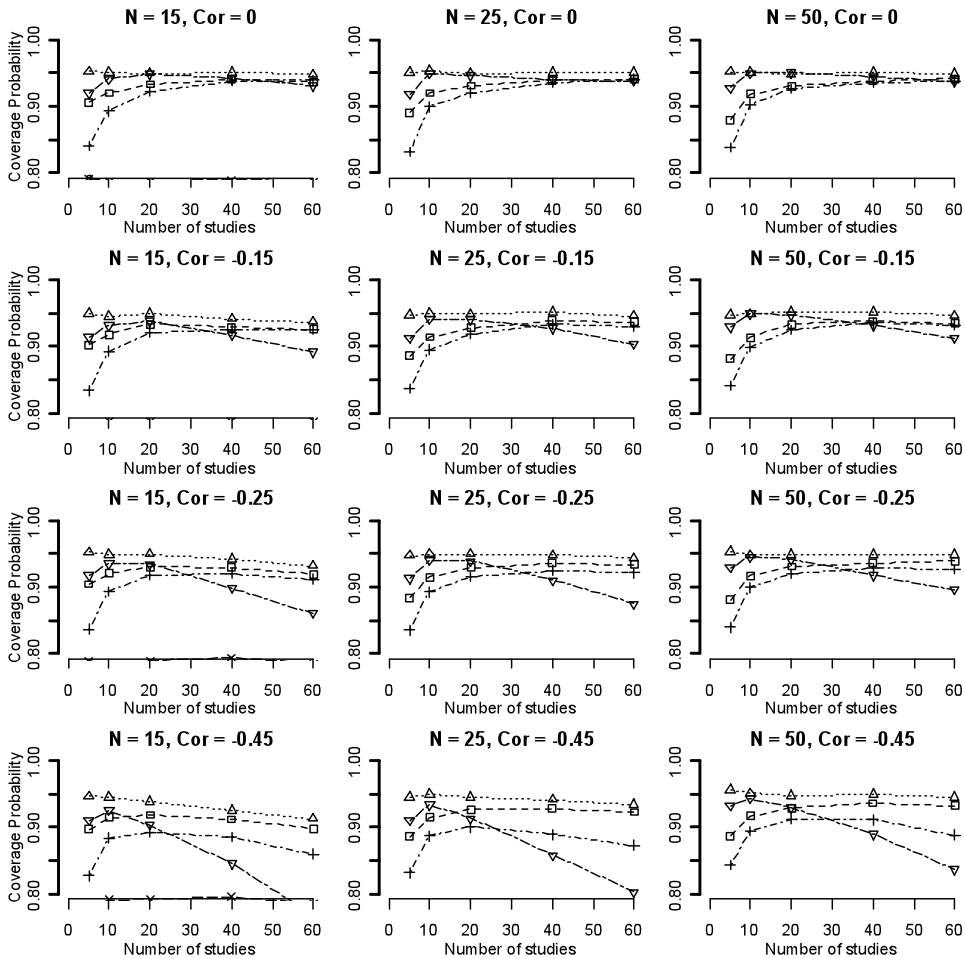


Figura 3. Cobertura de intervalo para distribución normal y $\tau^2 = 0{,}24$

Manteniendo la distribución normal subyacente pero con aumento de la variabilidad entre los efectos (Figura 3), se observó un mayor deterioro en el ajuste a la tasa nominal del intervalo de confianza conforme aumentaba la correlación entre tamaño muestral y efecto. Las tendencias fueron similares a las observadas en la Figura 2.

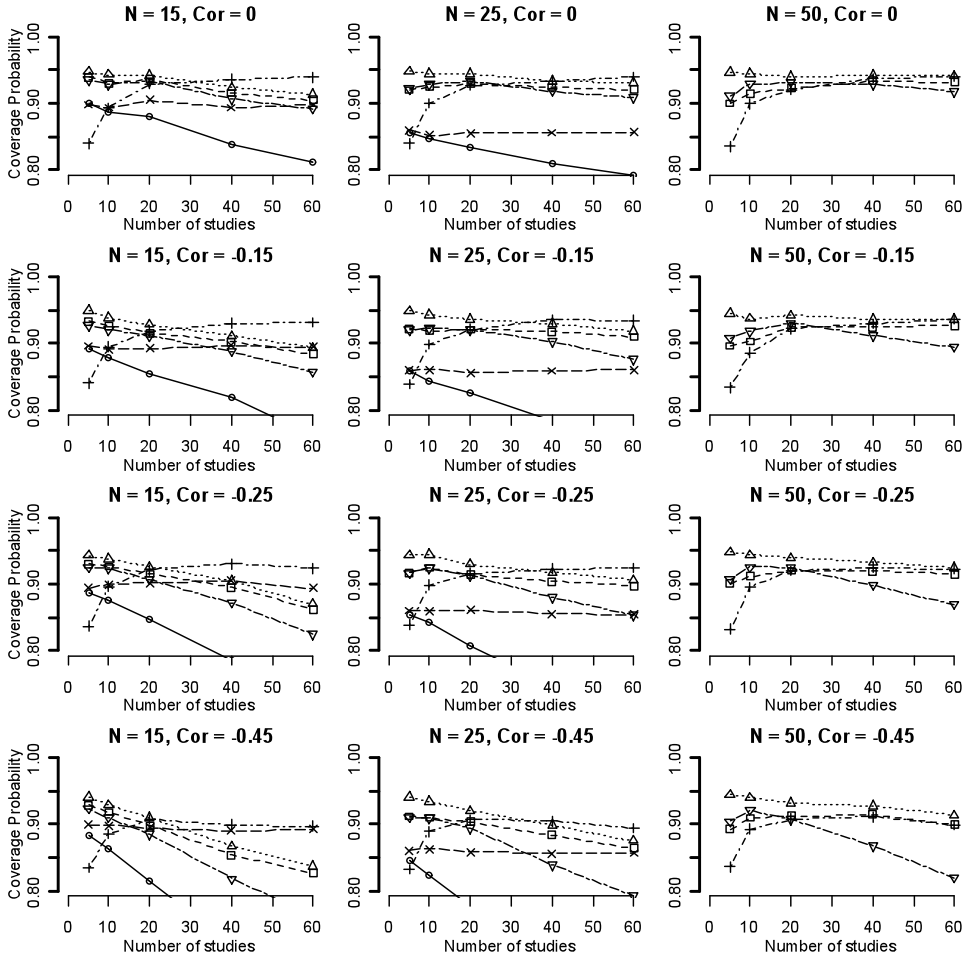


Figura 4. Cobertura de intervalo para distribución log-normal y $\tau^2 = 0{,}08$

En las condiciones con distribución log-normal subyacente y variabilidad moderada (Figura 4), el funcionamiento de los métodos fue similar al descrito en las Figuras 2 y 3, observándose una ligera mejora en el funcionamiento del método EA ponderando por el tamaño muestral.

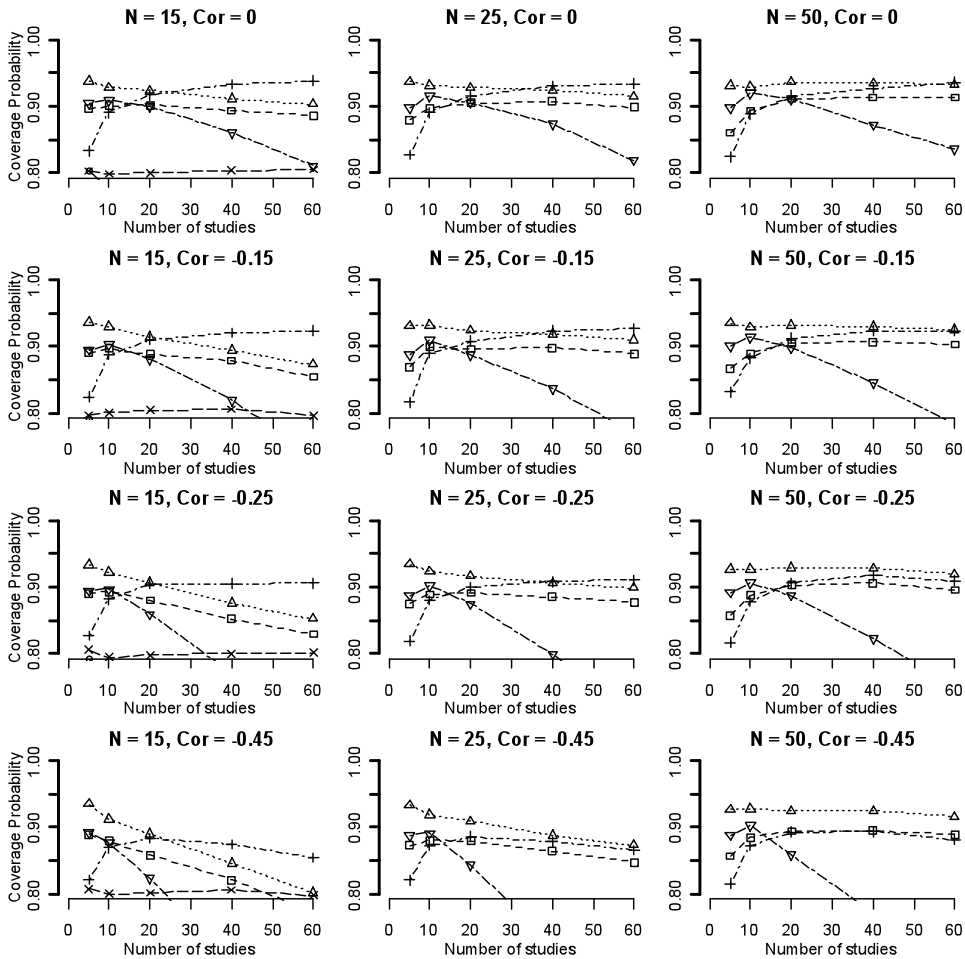


Figura 5. Cobertura de intervalo para distribución log-normal y $\tau^2 = 0{,}24$

Finalmente, la Figura 5 muestra un marcado deterioro de los métodos en las condiciones con distribución log-normal subyacente y alta variabilidad, especialmente a medida que se incrementó la correlación entre tamaño muestral y efecto. El método EA corregido por Hartung proporcionó una vez más el mejor ajuste de cobertura.

DISCUSIÓN

Este trabajo analizó, mediante simulación Monte Carlo, el funcionamiento de diversas alternativas coexistentes en la actualidad para realizar una estimación por intervalo del efecto medio en meta-análisis. La cobertura de los intervalos fue comparada en múltiples escenarios, en los que se trató de reflejar de manera realis-

ta una amplia gama de condiciones que pueden encontrarse al llevar a cabo un meta-análisis.

Tal y como se esperaba, asumir un modelo de efecto fijo resultó en una cobertura de intervalo apropiada sólo en las condiciones en las que existía un efecto paramétrico común. Las tasas de cobertura fueron, por el contrario, muy bajas cuando los efectos paramétricos eran heterogéneos. En estas condiciones, el modelo de coeficientes variables también mostró un marcado descenso de la cobertura del intervalo, lo cual se explica por el hecho de que este método no está específicamente diseñado para estimar más allá del simple promedio de los efectos paramétricos que estiman los respectivos estudios del meta-análisis.

Bonett (2009) y Shuster (2010) criticaron el uso generalizado del modelo EA en la actualidad, aduciendo ciertas amenazas cuyo influjo pusimos a prueba en este estudio de simulación. Por una parte, se asumió una distribución asimétrica (log-normal) para los efectos paramétricos, además de la habitual distribución normal, encontrándose un deterioro en el funcionamiento de casi todos los métodos. Este hallazgo encaja con las hipótesis previas, y nos lleva a aconsejar un análisis previo de la distribución de las estimaciones del tamaño del efecto (e.g., mediante un histograma), interpretando los resultados con cautela si la tendencia observada se aleja en exceso de la normalidad.

También se examinó el influjo de la existencia de una correlación negativa entre tamaño muestral y efecto. Tal y como Shuster (2010) advirtió, se produjo un deterioro en el funcionamiento de las técnicas que incluyen un factor de ponderación (en nuestro estudio, todas excepto el método de coeficientes variables). Los valores se alejaron considerablemente de la tasa nominal de cobertura cuando se combinó una correlación de magnitud elevada con una distribución asimétrica en los efectos. En estas condiciones, un incremento en el tamaño muestral medio condujo a un mayor equilibrio en los resultados. En concreto, cuando el tamaño muestral medio era de 50 sujetos, el método EA corregido por Hartung funcionó razonablemente bien con independencia del resto de condiciones. El incremento en el número de estudios, por otra parte, no conllevó una mejoría clara en los resultados para todos los métodos.

Los resultados encontrados aquí refrendan el uso de la corrección de Hartung (1999) cuando se asume un modelo EA en meta-análisis. En última instancia, la elección del modelo debería estar guiada por el margen de generalización que quiera alcanzar el investigador. Estrictamente, el proceso de muestreo de estudios en meta-análisis no es aleatorio, pero ésta es una limitación común a los estudios primarios de cualquier ámbito (e.g., Laird y Mosteller, 1990; Schmidt *et al.*, 2009). Dado que el objetivo de un meta-análisis suele ser obtener conclusiones válidas más allá de la muestra de estudios incluidos, nosotros defendemos el uso de un modelo EA siempre que se cumplan ciertas condiciones. En concreto, es necesario que el número de estudios no sea demasiado bajo, para que las estimaciones de los hiperparámetros sean precisas y representativas de la población a la que hacen referencia (Borenstein *et al.*, 2010). Otros aspectos, como la correlación entre tamaño muestral y efecto o la distribución subyacente de los efectos, deben ser debida-

mente evaluados, de manera que podamos determinar en qué medida son frecuentes en meta-análisis y, en caso de serlo, a qué pueden atribuirse. Como muestra este estudio, no obstante, la existencia de cualquiera de ellos aconseja una interpretación cautelosa de los resultados.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación (Proyecto nº PSI2009/12172).

REFERENCIAS

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods,* 14, 225-238.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., y Rothstein, H. R. (2010). A basic introduction to fixed-effects and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97-111.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal, 41,* 901-916.

Hedges, L. V., y Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Henmi, M., y Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, *29*, 2969-2983.

Hunter, J. E., y Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting errors and bias in research findings* (2ª ed.). Newbury Park, CA: Sage.

Laird, N. M., y Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6,* 5-30.

Sánchez-Meca, J., y Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*, 31-48.

Sánchez-Meca, J., Marín-Martínez, F., y López-López, J. A. (2011). Meta-análisis e intervención psicosocial basada en la evidencia. *Psychosocial Intervention, 20,* 95-107.

Schmidt, F. L., Oh, I.-S., y Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62*, 97-128.

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine, 29*, 1259-1265.

LA EVALUACIÓN DE LA CAPACIDAD PREDICTIVA EN LOS MODELOS DE META-REGRESIÓN

Fulgencio Marín-Martínez[1], José Antonio López-López[1] y Wolfgang Viechtbauer[2]
[1] Universidad de Murcia
[2] Universidad de Maastricht
Correo electrónico: fulmarin@um.es

Resumen

En un meta-análisis, los modelos de meta-regresión permiten analizar la posible asociación entre una o más variables moderadoras escogidas entre las características diferenciales de los estudios y los tamaños del efecto, evaluando tanto la significación estadística como la magnitud de la asociación. En el presente trabajo, a través de un meta-análisis real y un estudio de simulación Monte Carlo, comparamos el comportamiento estadístico de tres procedimientos para la estimación de la magnitud de la asociación entre un predictor y los tamaños del efecto en una meta-regresión: procedimientos de DerSimonian y Laird, máxima verosimilitud y máxima verosimilitud restringida. El procedimiento de DerSimonian y Laird mostró sistemáticamente el mejor comportamiento, con los menores valores de sesgo y media cuadrática de error.

El meta-análisis aporta una serie de procedimientos estadísticos para integrar cuantitativamente los resultados de múltiples estudios sobre un mismo tema. Una vez seleccionados los estudios de un determinado meta-análisis y expresados sus resultados en un mismo índice del tamaño del efecto (p. ej. la diferencia media tipificada o un coeficiente de correlación), un objetivo fundamental del meta-análisis es el de explicar las razones por las que tales resultados o efectos habitualmente van a diferir entre los estudios (Borenstein, Hedges, Higgins y Rothstein, 2009; Cooper, Hedges y Valentine, 2009; Sánchez-Meca y Marín-Martínez, 2010). En este contexto, los modelos de meta-regresión permiten evaluar la posible asociación entre ciertas características diferenciales de los estudios (p. ej. el año de publicación o la edad media de la muestra) y los valores del tamaño del efecto en cada estudio, dando cuenta de algunos de los factores que podrían explicar parte de la heterogeneidad de los resultados en los estudios de un meta-análisis.

Cuando el propósito de un meta-análisis es el de generalizar sus resultados más allá de la muestra de estudios incluidos en la revisión, se aconseja comenzar asumiendo un modelo de efectos aleatorios, donde la variabilidad de los tamaños

del efecto se descompone en la varianza inter-estudios, τ^2_δ, y la varianza intra-estudios, v_i (Borenstein *et al.*, 2009; Raudenbush, 2009; Sánchez-Meca y Marín-Martínez, 2008). En un meta-análisis de *k* estudios, donde el resultado del estudio *i* se estima como una diferencia media tipificada, d_i, siendo δ_i el tamaño del efecto real en la población y $e_i = d_i - \delta_i$ el error muestral de estimación, el modelo de efectos aleatorios se formula según:

$$d_i = \beta_0 + u_i + e_i, \quad (1)$$

donde β_0 representa el tamaño del efecto medio en la distribución de efectos paramétricos (la media de los valores δ_i) y $u_i = \delta_i - \beta_0$ la desviación entre el efecto paramétrico de cada estudio y la media de los efectos paramétricos. Habitualmente se asume que los errores e_i y u_i se distribuyen normal e independientemente, según $e_i \sim N(0, v_i)$, donde v_i es la varianza intra-estudios o debida al error de muestreo en la elección de las unidades de cada estudio, y $u_i \sim N(0, \tau^2_\delta)$, siendo τ^2_δ la varianza inter-estudios o debida a las características diferenciales de los estudios.

La incorporación en el modelo (1) de una o más covariables o predictores escogidos entre las características diferenciales de los estudios, con el objetivo de explicar parte de la variabilidad inter-estudios, nos permite definir un modelo de meta-regresión, que en el caso más sencillo con un solo predictor se formula según:

$$d_i = \beta_0 + \beta_1 X_i + u_i + e_i, \quad (2)$$

siendo X_i el valor del predictor *X* en el estudio *i* del meta-análisis y β_1 la pendiente de la meta-regresión. A diferencia de lo que ocurría en el modelo (1), en este caso los errores u_i se distribuyen según $u_i \sim N(0, \tau^2_{\delta/X})$, siendo $\tau^2_{\delta/X}$ la varianza inter-estudios residual o la parte de la varianza inter-estudios que no ha podido ser explicada por el predictor.

A partir de la comparación entre los modelos (1) y (2), o más concretamente entre los valores estimados de la varianza inter-estudios total en un modelo sin predictores, $\hat{\tau}^2_\delta$, y de la varianza inter-estudios residual en el mismo modelo con al menos un predictor, $\hat{\tau}^2_{\delta/X}$, se define el índice R^2 o de proporción de varianza explicada por el predictor según (Borenstein *et al.*, 2009; Raudenbush, 2009):

$$R^2 = \frac{\hat{\tau}^2_\delta \text{ (modelo sin predictores)} - \hat{\tau}^2_{\delta/X} \text{ (modelo con el predictor } X)}{\hat{\tau}^2_\delta \text{ (modelo sin predictores)}}, \quad (3)$$

que permite cuantificar la capacidad predictiva del modelo de meta-regresión como el porcentaje de la varianza inter-estudios que conseguimos explicar a partir del predictor o predictores introducidos en el modelo.

Existen múltiples procedimientos para estimar la varianza inter-estudios de un meta-análisis cuyos resultados no suelen coincidir (Raudenbush, 2009; Viechtbauer, 2005). Sustituyendo en la ecuación (3) por los diferentes estimadores de τ^2_δ y $\tau^2_{\delta/X}$, cabe esperar que los resultados de R^2 también varíen. Centrándonos en los tres procedimientos más habituales de estimación de la varianza inter-estudios:

DerSimonian y Laird, máxima verosimilitud y máxima verosimilitud restringida (Raudenbush, 2009, ecuaciones 16.43, 16.32 y 16.37), el objetivo del presente trabajo es el de evaluar las consecuencias del empleo de diferentes estimadores de la varianza inter-estudios en la correcta estimación de R^2 o de la capacidad predictiva de una meta-regresión.

Ejemplo de un meta-análisis real

En un meta-análisis sobre las diferencias de género en rendimiento académico tomado del informe PISA del año 2003 (Else-Quest, Hyde & Linn, 2010), se escogió como índice del tamaño del efecto la diferencia media tipificada, *d*, donde el signo positivo indicaba el mayor rendimiento de los varones. El meta-análisis integraba 18 estudios realizados cada uno de ellos en un país diferente, con resultados significativamente heterogéneos, oscilando entre el valor $d = -0{,}17$ de Islandia y $d = +0{,}25$ de Corea del Sur.

Para explicar esta heterogeneidad de resultados en los estudios se aplicó una meta-regresión con un predictor, el porcentaje de escaños ocupados por mujeres en los Parlamentos de cada país, encontrándose una asociación significativa en el sentido de que a mayor presencia femenina en los Parlamentos menores eran las diferencias de género. Estimado el porcentaje de varianza explicada a partir del índice R^2 (ecuación 3) con los tres procedimientos de estimación de la varianza inter-estudios de DerSimonian y Laird (DL), máxima verosimilitud (MV) y máxima verosimilitud restringida (MVR), se encontraron diferentes valores que oscilaban entre el 3,29% y el 21,91%: $R^2_{DL} = 3{,}29\%$, $R^2_{MVR} = 16{,}29\%$ y $R^2_{MV} = 21{,}91\%$.

Con el propósito de indagar sobre cuál de estos tres estimadores sería el más adecuado en un determinado meta-análisis, hicimos un estudio de simulación Monte Carlo sobre su comportamiento estadístico en función de las características del meta-análisis.

Un estudio Monte Carlo

Con el programa Gauss (Aptech Systems, 2001) se simularon meta-análisis de *k* estudios donde el índice del tamaño del efecto era la diferencia media tipificada. Para cada meta-análisis se generaron dos vectores, $\boldsymbol{\delta}$ y, ambos de dimensión *k x1*, que simulaban los tamaños del efecto paramétricos y los valores de un predictor cuantitativo, respectivamente. Los valores de $\mathbf{X}$ se generaron a partir de una distribución $N(0,1)$ y los de $\boldsymbol{\delta}$ a partir de la expresión $\boldsymbol{\delta} = \beta_1\mathbf{X} + \mathbf{u}$, donde la pendiente β_1 era un valor constante y el término de error $\mathbf{u}$ se distribuía según $N(0{,}5; \tau^2_{\delta/X})$, siendo $\tau^2_{\delta/X}$ la varianza inter-estudios residual. Así, el vector $\boldsymbol{\delta}$ se distribuyó según $N(0{,}5; \tau^2_{\delta})$, siendo τ^2_{δ} la varianza inter-estudios total. La Tabla 1 presenta las siete combinaciones de valores paramétricos de la varianza inter-estudios total y el porcentaje de varianza explicada por el predictor manipuladas en la simulación,

con las que se obtuvieron los correspondientes valores de β_1 y $\tau^2_{\delta/X}$ a partir de la relación $\tau^2_{\delta} = \beta_1^2 + \tau^2_{\delta/X}$.

Tabla 1. Valores paramétricos de la varianza inter-estudios (τ^2_{δ}) y el porcentaje de varianza explicada por el predictor (ρ^2) manipulados en la simulación

τ^2_{δ}	0	0,08			0,32		
ρ^2	0%	0%	25%	50%	0%	25%	50%

Otros factores manipulados fueron el número de estudios del meta-análisis (k = 10, 20, 40) y el tamaño muestral medio de los estudios del meta-análisis ($\overline{N}$ = 30, 50, 100), siendo 63 (7x3x3) el total de condiciones manipuladas y 10.000 el número de replicas por condición. Cada vez que se generaba un efecto paramétrico δ_i procedente del vector **δ**, se obtenía el estudio *i* de un meta-análisis, con dos grupos independientes procedentes de las poblaciones $N(\delta_i, 1)$ y $N(0, 1)$. En cada estudio de un meta-análisis se calculó la diferencia media tipificada (Marín-Martínez y Sánchez-Meca, 2010, ecuación 2), y a través de los *k* estudios de cada meta-análisis se aplicaron los tres procedimientos de DerSimonian y Laird (DL), máxima verosimilitud (MV) y máxima verosimilitud restringida (MVR) (Raudenbush, 2009, ecuaciones 16.43, 16.32 y 16.37) para estimar la varianza inter-estudios total, la varianza inter-estudios residual y la proporción de la varianza inter-estudios explicada por el predictor (ecuación 3). Finalmente, a través de las 10.000 réplicas de cada meta-análisis se estimó el sesgo y la media cuadrática de error de todos estos estimadores.

RESULTADOS Y DISCUSIÓN

La Tabla 2 presenta los valores medios del sesgo y la media cuadrática de error (MCE), a través de las 63 condiciones manipuladas, de los tres procedimientos DL, MV y MVR como estimadores de la varianza inter-estudios total, τ^2_{δ}, la varianza inter-estudios residual, $\tau^2_{\delta/X}$ y la proporción de varianza explicada en una meta-regresión, ρ^2. Los resultados relativos a los estimadores de τ^2_{δ} eran los esperados al coincidir con los de otros estudios previos (Raudenbush, 2009; Viechtbauer, 2005): en promedio el estimador $\hat{\tau}^2_{MV}$ figura como el más preciso o con la menor MCE (0,010), aunque a costa de un importante sesgo negativo (-0,025), y el estimador $\hat{\tau}^2_{MVR}$ corrige tal sesgo negativo reduciéndolo a -0,010, aunque a costa de presentar la mayor MCE (0,012).

Sin embargo, lo que no esperábamos es que este patrón de resultados cambiase de forma importante cuando aplicados los mismos procedimientos en el contexto de una meta-regresión, estiman la varianza la varianza inter-estudios residual, $\tau^2_{\delta/X}$. Como se observa en la Tabla 2, los tres procedimientos incrementan tanto su sesgo negativo promedio como su MCE promedio, manteniéndose el orden de los procedimientos en cuanto al grado de sesgo pero no en cuanto al grado de MCE: mientras que MV es el procedimiento con la menor MCE en la estimación de τ^2_{δ} (0,010), se convierte en el procedimiento con la mayor MCE en la estimación de $\tau^2_{\delta/X}$ (0,028)

y DL es el procedimiento más preciso o con la menor MCE en la estimación de $\tau^2_{\delta/X}$ (0,014).

Puesto que la proporción de varianza explicada en una meta-regresión, R^2, es función directa de los valores estimados de la varianza inter-estudios total y residual (ver ecuación 3), el comportamiento que acabamos de ver de los tres procedimientos DL, MV y MVR como estimadores de τ^2_{δ} y $\tau^2_{\delta/X}$ nos ayuda a entender los resultados relativos al objetivo fundamental de este trabajo: las propiedades estadísticas de DL, MV y MVR como estimadores de ρ^2. Tal y como se observa en la Tabla 2, los tres procedimientos muestran un sesgo promedio positivo de 0,013 para DL, 0,020 para MVR y 0,089 para MV, lo que concuerda con el conocido sesgo positivo del índice de proporción de varianza explicada, R^2, en un análisis de regresión convencional. Comparando los tres procedimientos nos encontramos con que DL es el mejor al presentar los menores promedios de sesgo (0,013) y MCE (0,063), seguido de cerca por MVR (sesgo = 0,020, MCE = 0,067) y a mayor distancia por MV (sesgo = 0,089, MCE = 0,087), con el peor comportamiento.

Tabla 2. Valores medios del sesgo y la media cuadrática de error (entre paréntesis) de los procedimientos de DerSimonian y Laird (DL), máxima verosimilitud (MV) y máxima verosimilitud restringida (MVR) como estimadores de la varianza inter-estudios total (τ^2_{δ}), la varianza inter-estudios residual ($\tau^2_{\delta/X}$) y la proporción de varianza explicada en una meta-regresión (ρ^2)

$\hat{\tau}^2_{DL}$	$\hat{\tau}^2_{MV}$	$\hat{\tau}^2_{MVR}$
-0,014	-0,025	-0,010
(0,011)	(0,010)	(0,012)
$\hat{\tau}^2_{DL/X}$	$\hat{\tau}^2_{MV/X}$	$\hat{\tau}^2_{MVR/X}$
-0,050	-0,074	-0,049
(0,014)	(0,028)	(0,015)
R^2_{DL}	R^2_{MV}	R^2_{MVR}
0,013	0,089	0,020
(0,063)	(0,087)	(0,067)

Las Figuras *1a, 1b* y *1c* muestran los valores MCE promedio de los tres procedimientos DL, MV y MVR como estimadores de ρ^2, en función del número de estudios del meta-análisis, el tamaño muestral medio de los estudios y del valor paramétrico del porcentaje de varianza explicada por el predictor, respectivamente. En las tres Figuras se observa cómo sistemáticamente y a través de todas las condiciones, el procedimiento DL se mostró como el más preciso (con la menor MCE), seguido muy de cerca por MVR y a más distancia por MV. Finalmente, la Figura *1c* muestra que las discrepancias entre los tres procedimientos tienden a desaparecer conforme aumenta el grado de asociación entre el predictor y los tamaños del efecto.

Figura 1a

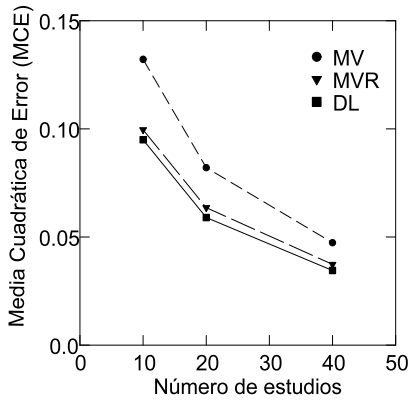


Figura 1b

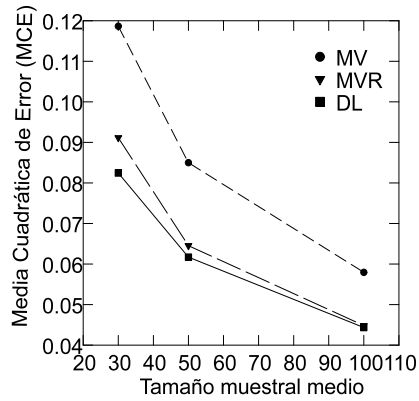


Figura 1c

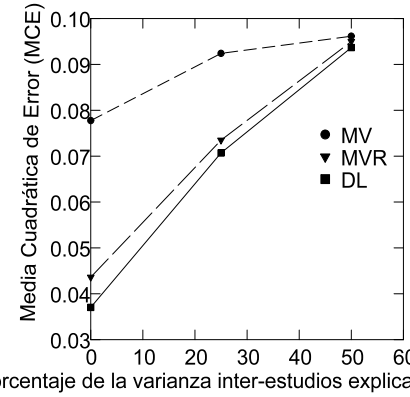


Figura 1. Valores promedio de la media cuadrática de error de los procedimientos de DerSimonian y Laird (DL), máxima verosimilitud (MV) y máxima verosimilitud restringida (MVR) como estimadores de la proporción de varianza explicada, en función del número de estudios del meta-análisis (1a), del tamaño muestral medio de los estudios (1b) y del porcentaje real de varianza explicada en una meta-regresión (1c).

CONCLUSIONES

Los tres procedimientos DL, MV y MVR habitualmente van a ofrecer resultados diferentes en la estimación de la proporción de varianza explicada en una meta-regresión, salvo en aquellos meta-análisis donde haya una fuerte asociación entre el predictor y los tamaños del efecto. A través de todas las condiciones manipuladas en la simulación, el procedimiento DL presentó el mejor comportamiento,

con el menor sesgo y la menor MCE, seguido muy de cerca por MVR y a más distancia por MV, que mostró los mayores valores de sesgo y MCE.

En la práctica se recomienda el empleo del procedimiento DL, sobre todo en los meta-análisis con pocos estudios, tamaños muestrales bajos y con un bajo o moderado grado de relación entre los tamaños del efecto y el predictor, que es donde se esperan las mayores discrepancias entre los valores de la proporción de varianza explicada estimados con los tres procedimientos.

Finalmente, las conclusiones de este trabajo se limitan a los meta-análisis donde el índice del tamaño del efecto es la diferencia media tipificada y a las condiciones específicas manipuladas en la simulación. Futuros estudios deberían explorar el comportamiento de estos procedimientos en meta-análisis con otros índices del tamaño del efecto, un número variable de predictores y otras condiciones relativas a la distribución de tamaños el efecto en la población y al valor paramétrico del grado de asociación entre los predictores y los tamaños del efecto.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación (Proyecto nº PSI2009/12172).

REFERENCIAS

Aptech Systems. (2001). *The GAUSS Program (Version 3.6)* [Software informático]. Kent, WA: Author.

Borenstein, M.J., Hedges, L.V., Higgins, J.P.T. y Rothstein, H.R. (2009). *Introduction to Meta-Analysis.* Chichester, UK: Wiley.

Cooper, H., Hedges, L.V. y Valentine, J.C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2ª ed.). Nueva York: Russell Sage Foundation.

Else-Quest, N.M., Hyde, J.S. y Linn, M.C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136,* 103-127.

Marín-Martínez, F. y Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement, 70,* 56-73.

Raudenbush, S.W. (2009). Random effects model. En H. Cooper, L.V. Hedges y J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295-315). Nueva York: Russell Sage Foundation.

Sánchez-Meca, J. y Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13,* 31-48.

Sánchez-Meca, J. y Marín-Martínez, F. (2010). Meta-analysis. En P. Peterson, E. Baker y B. McGaw (Eds.), *International Encyclopedia of Education*, Volumen 7 (3ª ed.) (pp. 274-282). Oxford: Elsevier.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30,* 261-293.

VALIDEZ Y FIABILIDAD DE UNA ESCALA PARA LA EVALUACIÓN DE LA CALIDAD DE LOS ESTUDIOS PRIMARIOS EN META-ANÁLISIS

José Antonio López-Pina[1], Julio Sánchez Meca[1] y Rosa Mª Núñez Núñez[2]

[1]Universidad de Murcia

[2]Universidad Miguel Hernández

Correo electrónico: jlpina@um.es

Resumen

En meta-análisis la valoración de la calidad metodológica de los estudios primarios es una tarea imprescindible para detectar posibles sesgos en las estimaciones de los efectos. Sin embargo, no existe un consenso en la comunidad científica sobre cómo debe evaluarse la calidad de los estudios primarios, dado que en la literatura experimental se han propuesto casi 200 escalas con muy diferentes formatos, la mayoría de ellas propuestas en el ámbito médico. En el ámbito de las ciencias del comportamiento, los instrumentos propuestos son escasos y apenas se han sometido a un análisis psicométrico de sus propiedades. En esta comunicación se presenta un estudio de validez y fiabilidad de una escala compuesta por 10 ítems dicotómicos elaborada por nuestro equipo de investigación para valorar la calidad metodológica de estudios primarios y ser utilizada en meta-análisis dirigidos a evaluar la eficacia de intervenciones psicológicas. Para examinar su validez y fiabilidad se aplicó ésta a los 66 estudios primarios de un meta-análisis sobre la eficacia de los tratamientos psicológicos del trastorno de pánico con o sin agorafobia. Se presentan los resultados del análisis de la validez de constructo y del grado de acuerdo inter-codificadores logrado en su aplicación.

En meta-análisis los estudios primarios se someten a un proceso de codificación en el que se registran las características sustantivas y metodológicas que pudieran estar estadísticamente relacionadas con las estimaciones de los efectos (Botella y Gambara, 2002; Littell, Corcoran y Pillai, 2008; Sánchez-Meca, 2008). En este contexto, una de las etapas fundamentales del meta-análisis consiste en evaluar la calidad metodológica de cada estudio para poder comprobar posteriormente el posible influjo de sesgos en las estimaciones de los efectos debidos a la falta de control de variables.

Las escalas de calidad reúnen múltiples ítems para evaluar la calidad de los estudios primarios, con un esquema de puntuación para cada ítem y posibilitan

atribuir a cada estudio una puntuación global de calidad que considere conjuntamente todos los aspectos que inciden en ella. Las escalas cuantitativas de calidad tienen como mayor ventaja el que, si están bien construidas y poseen una aceptable validez y fiabilidad, aportan una información muy útil sobre el nivel global de calidad en un estudio, resumiendo en un único indicador toda la información.

En el ámbito médico se han desarrollado múltiples escalas de calidad, de entre las que destacan la de Chalmers, Smith, Blackburn et al. (1981) cuyos ítems valoran componentes de la validez interna, validez externa, presentación de datos, análisis estadístico y organización de un ensayo clínico aleatorizado (ECA); y la escala de Jadad, Moore, Carroll et al. (1996), que se centra exclusivamente en tres dimensiones de validez interna: aleatorización, enmascaramiento (*blinding*) y pérdida de sujetos, pero da más peso a la calidad del reporte que a la calidad metodológica real. Otras guías para evaluar la calidad son: la guía CONSORT (Moher, Jones y Lepage for the CONSORT Group, 2001), desarrollada para mejorar la calidad del reporte de los ECAs y su versión en ciencias del comportamiento (Boutron et al. for the CONSORT Group, 2008), y la guía TREND (Des Jarlais, Lyles, Crepaz and the TREND Group, 2004) dirigida a mejorar el reporte de los estudios evaluativos no aleatorizados también en salud pública.

En el ámbito de las Ciencias del Comportamiento, Miller y Wilbourne (2002) elaboraron la *Methodological Quality Rating Scale* (MQRS), que incluye 12 ítems sobre diseño del estudio, mortalidad, duración del seguimiento, tipos de medidas de resultado y control de la calidad de la intervención. En el ámbito de la criminología, Aschcroft, Daniels y Flores (2004) aplicaron una escala de calidad que tomaba en consideración aspectos tales como el tipo de diseño (aleatorizado versus no aleatorizado), el tamaño muestral, la mortalidad experimental y el uso de instrumentos de medida válidos y fiables. Por último, Valentine y Cooper (2008) desarrollaron el *Design and Implementation Assessment Device* (DIAD), que propone un sistema jerárquico de identificación de dimensiones de calidad en el que se incluyen 32-34 ítems que permiten operacionalizar la valoración de la calidad del diseño e implementación de los estudios evaluativos.

La mayoría de las escalas de calidad propuestas en la literatura no se han sometido a un análisis psicométrico y su ámbito de aplicación se centra básicamente en ciencias de la salud, sobre todo en medicina. Son escasos los esfuerzos que se han hecho para proponer una escala de calidad de fácil aplicación en ciencias del comportamiento para valorar la calidad metodológica de estudios empíricos que evalúan la eficacia de programas (psicológicos, sociales, educativos, criminológicos, etc.). Tomando como base la experiencia de nuestro equipo de investigación en la realización de meta-análisis sobre la eficacia de intervenciones, y una vez hecha una revisión de la literatura, elaboramos una escala compuesta por 10 ítems dicotómicos (verdadero/falso) que permitieran valorar la calidad metodológica de los estudios empíricos en los que el diseño implica la asignación de sujetos a una o varias condiciones de tratamiento, pudiéndose aplicar en el contexto general de las ciencias sociales. El propósito de esta investigación fue analizar la fiabilidad

inter-observadores de nuestra escala de calidad. Para ello, utilizamos la metodología del estudio de caso, mediante la aplicación de la escala a los estudios de un meta-análisis realizado por nuestro equipo de investigación.

MÉTODO

Muestra

Para analizar el grado de acuerdo inter-observadores del proceso de codificación de los ítems de nuestra escala de calidad metodológica, aplicamos ésta a los 66 estudios de un meta-análisis sobre la eficacia de los tratamientos psicológicos del trastorno de pánico con o sin agorafobia (Sánchez-Meca, Rosa-Alcázar, Marín-Martínez y Gómez-Conesa, 2010). Aunque para la aplicación de nuestra escala de calidad no era obligatorio que el diseño incluya un grupo de control, los 66 estudios del meta-análisis se caracterizaban por incluir todos ellos un grupo de control.

Instrumento

La escala está formada por diez ítems con formato dicotómico (1, presente; 0, ausente) que valoran diversos aspectos relacionados con la calidad metodológica del estudio evaluativo: (1) asignación aleatoria de los sujetos, (2) igualación de los grupos en el pretest en las variables dependientes y en las variables de control (al menos en el 90% de las variables), (3) reporte de los resultados en el pretest para al menos el 90% de las variables dependientes, (4) inclusión de un grupo de control con placebo psicológico, (5) inclusión de un grupo de control con placebo farmacológico, (6) enmascaramiento del evaluador, (7) uso de instrumentos debidamente validados, (8) tamaño muestral total en el postest por encima de la media (en nuestro caso, $N \geq 34$), (9) mortalidad diferencial en el postest entre los grupos tratado y de control igual al 10% como máximo, y (10) uso de análisis por intención de tratar.

Procedimiento

Se crearon dos equipos de codificadores formados por dos codificadores cada uno. Tras el debido entrenamiento de los codificadores, se realizó un estudio piloto con una muestra aleatoria de 10 estudios a los que fue aplicada la escala de calidad. Las discrepancias encontradas entre los codificadores fueron resueltas con la finalidad de construir un manual de codificación claro y conciso sobre las valoraciones dicotómicas de los ítems de la escala. Posteriormente, los dos equipos de codificadores codificaron los 66 estudios del meta-análisis para valorar la fiabilidad inter-observadores mediante el cálculo del coeficiente *kappa* de Cohen para cada ítem de la escala.

RESULTADOS

La Tabla 1 presenta los coeficientes kappa de Cohen obtenidos con cada ítem y la frecuencia de ocurrencia de cada uno de los ítems de la escala de calidad. Los ítems 3 y 7 obtuvieron frecuencias del 100%, dado que ambos fueron utilizados como criterios de selección de los estudios para realizar el meta-análisis sobre la eficacia de los tratamientos psicológicos del trastorno de pánico con o sin agorafobia, por lo que no se pudo determinar el coeficiente de acuerdo entre codificadores. El coeficiente de acuerdo más elevado se encontró en el ítem sobre el *Tamaño muestral en el postest* (0,970), seguido por la existencia de un *Grupo control placebo farmacológico* (0,868), *Grupo control placebo psicológico* (0,816), *Enmascaramiento del evaluador* (0,791), *Análisis por intención de tratar* (0,774), *Mortalidad diferencial en el postest* (0,708), *Asignación aleatoria de los sujetos* (0,663) e *Igualación del los grupos en el pretest en la VDs y variables control* (0,486).

Tabla 1. Acuerdo inter-jueces (con sus errores típicos) y frecuencia (porcentaje) de ocurrencia de los ítems de la escala

Ítems	Acuerdo Inter-jueces	Frecuencia	
	Kappa	0	1
1. Asignación aleatoria de los sujetos	0,663 (0,140)	8 (12,1%)	58 (87,9%)
2. Igualación de los grupos en el pretest en las VDs y variables de control (≥ 90%)	0,486 (0,156)	9 (13,6%)	57 (86,4%)
3. Inclusión de las medidas pretest en al menos el 90% de las VDs	–	0 (0%)	66 (100%)
4. Inclusión de un grupo de control placebo psicológico	0,816 (0,103)	57 (86,4%)	9 (13,6%)
5. Inclusión de un grupo de control placebo farmacológico	0,868 (0,074)	52 (78,8%)	14 (81,2%)
6. Enmascaramiento del evaluador	0,791 (0,081)	44 (66,7%)	22 (33,3%)
7. Uso de instrumentos debidamente validados	–	0 (0%)	66 (100%)
8. Tamaño muestral total en el postest ($N \geq 34$)	0,970 (0,030)	33 (50%)	33 (50%)
9. Mortalidad diferencial en el postest ≤ 10%	0,708 (0,090)	39 (59,1%)	27 (40,9%)
10. Análisis por intención de tratar	0,774 (0,080)	41 (62,1%)	25 (37,9%)

En general, la presencia de los ítems de calidad seleccionados para esta escala de calidad fue elevada, aunque desigual. Así, además de los ítems 3 y 7 ya mencionados arriba, el mayor porcentaje de frecuencia se obtuvo en el ítem de *Asignación aleatoria de los sujetos* (87,9%), seguido del de *Igualación de los grupos en el*

pretest en las VDs y variables control (86,4%), *Grupo control con placebo farmacológico* (81,2%), *Tamaño muestral total en postest* (50%), *Mortalidad diferencial en el postest* (40,9%), *Análisis por intención de tratar* (37,9%), *Enmascaramiento del evaluador* (33,3%) y *Grupo control con placebo psicológico* (13,6%).

CONCLUSIONES

El propósito de esta investigación fue presentar una escala elaborada por nuestro equipo de investigación para valorar la calidad metodológica de los estudios primarios que evalúan la eficacia de programas en el ámbito de las ciencias sociales, para poder examinar la existencia de sesgos en las estimaciones de los efectos cuando se lleva a cabo un meta-análisis que integra un determinado conjunto de estudios evaluativos sobre un problema común. Los ítems que componen la escala se elaboraron tomando como base una revisión de la literatura, así como la experiencia de nuestro equipo de investigación realizando meta-análisis sobre la eficacia de intervenciones. El examen de los ítems pone de manifiesto su similitud con muchos de los ítems que habitualmente componen otras escalas de calidad propuestas en la literatura. El objetivo fundamental de nuestro estudio consistió en analizar la fiabilidad inter-observadores que nuestra escala es capaz de alcanzar cuando se aplica a un conjunto de estudios primarios de un meta-análisis. No debemos olvidar que, dependiendo del grado de claridad y detalle con que los estudios empíricos reportan sus características metodológicas, el proceso de codificación de tales características puede resultar más o menos fácil. Es por ello que nuestro propósito era construir una escala con las suficientes indicaciones y consignas como para lograr un alto grado de acuerdo inter-jueces.

El alto grado de acuerdo alcanzado entre los codificadores pone de manifiesto la excelente fiabilidad inter-observadores de nuestra escala de calidad. No obstante, es preciso tener en cuenta que los resultados obtenidos proceden de una aplicación concreta de la escala a un meta-análisis. El grado de acuerdo inter-codificadores puede variar dependiendo de la variabilidad metodológica exhibida por un conjunto de estudios primarios, de forma que cuanto mayor sea dicha variabilidad tanto mayores serán las estimaciones de la fiabilidad inter-observadores alcanzadas por los ítems de la escala. El conjunto de estudios utilizados en nuestra investigación se caracterizaban por tener una buena calidad metodológica en general, ya que todos ellos incluían medidas pretest y un grupo de control. Es de esperar que nuestra escala de calidad alcance incluso mejores resultados si se aplica a una base de estudios más diversos en cuanto a los tipos de diseños utilizados.

La investigación futura en relación a nuestra escala de calidad debe pasar por una depuración de los ítems para lograr una mayor validez de constructo y su comparación con otras escalas de calidad propuestas en la literatura para examinar su validez convergente. En esta misma línea, el estudio comparativo de varias escalas de calidad aplicadas a una misma base de estudios permitirá comprobar en qué grado dichas escalas alcanzan resultados similares o, como ya se ha compro-

bado en estudios previos dentro del ámbito médico, se observan importantes discrepancias entre ellas (Herbison, Hay-Smith y Gillespie, 2006).

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación (Proyecto nº PSI2009/12172).

REFERENCIAS

Ashcroft, J., Daniels, D.J. y Flores, J.R. (2004). *Blueprints for Violence Prevention*. Washington, DC: U.S. Department of Justice.

http://www.ncjrs.org/pdffiles1/ojjdp/204274.pdf

Botella, J. y Gambara, H. (2002). *Qué es el meta-análisis*. Madrid: Biblioteca Nueva.

Boutron, I., Moher, D., Altman, D.G., Schulz, K.F. y Ravaud, P., for the CONSORT Group (2008). Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration. *Annals of Internal Medicine, 148*, 295-309.

Chalmers, T.C., Smith, H., Balckburn, B., et al. (1981). A method for assessing the quality of a randomized controlled trial. *Controlled Clinical Trials, 2*, 31-49.

Des Jarlais, D.C., Lyles, C., Crepaz, N. and the TREND Group (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94*, 361-366.

Herbison, P., Hay-Smith, J. y Gillespie, W.J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology, 59*, 1249-1256.

Jadad, A.R., Moore, R.A., Carroll, D. et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*, 1-12.

Littell, J.H., Corcoran, J. y Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford, UK: Oxford University Press.

Miller, W. y Wilbourne, P. (2002). Mesa Grande: A methodological analysis of clinical trials of treatments for alcohol use disorders. *Addiction, 97*, 265-277.

Moher, D., Jones, A. y Lepage, L. for the CONSORT Group (2001). Use of the CONSORT statement and quality of reports for randomized trials: A comparative before-and-after evaluation. *Journal of the American Medical Association, 285*, 1992-1995.

Sánchez-Meca, J. (2008). Meta-análisis de la investigación. En M.A. Verdugo, M. Crespo, M. Badía y B. Arias (Coords.), *Metodología en la investigación sobre discapacidad: Introducción al uso de las ecuaciones estructurales* (pp. 121-139). Salamanca: Publicaciones del INICO.

Sánchez-Meca, J., Rosa-Alcázar, A.I., Marín-Martínez, F. y Gómez-Conesa, A. (2010). The psychological treatment of panic disorder with and without agoraphobia: A meta-analysis. *Clinical Psychology Review*, *30*, 37-50.

Valentine, J.C. y Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*, 130-149.

VALORACIÓN DE LA CALIDAD DE LOS ESTUDIOS PRIMARIOS EN META-ANÁLISIS Y SU RELACIÓN CON EL TAMAÑO DEL EFECTO

Julio Sánchez Meca, José A. López Pina y Fulgencio Marín Martínez

Universidad de Murcia
Correo electrónico: jsmeca@um.es

Resumen

Uno de los objetivos prioritarios en meta-análisis es valorar la calidad metodológica de los estudios primarios con objeto de comprobar si ésta puede estar estadísticamente relacionada con los tamaños del efecto. Si existe una relación estadística entre calidad del estudio y tamaño del efecto, entonces ésta es evidencia de que los estudios con una pobre calidad pueden estar ofreciendo estimaciones sesgadas de los efectos paramétricos. El objetivo de esta investigación es demostrar, mediante una aplicación a un meta-análisis real, cómo determinados aspectos relacionados con la calidad metodológica de los estudios primarios pueden estar estadísticamente relacionados con los tamaños del efecto obtenidos en dichos estudios. Para ello, hemos aplicado una escala de calidad compuesta por diez ítems dicotómicos y elaborada por nuestro equipo de investigación a un meta-análisis concreto. Nuestra hipótesis es que cuando un estudio primario no cumple con los criterios de calidad definidos en cada uno de estos ítems, la estimación del tamaño del efecto ofrecida por dicho estudio exhibirá un sesgo positivo del verdadero efecto en la población y, en consecuencia, la puntuación total de la escala de calidad estará negativamente relacionada con el tamaño del efecto. Finalmente, se discuten las implicaciones de nuestros resultados.

Cuando se lleva a cabo un meta-análisis, uno de los objetivos prioritarios es examinar el influjo de variables moderadoras y características de los estudios sobre la variabilidad de los tamaños del efecto obtenidos en los estudios primarios. En un buen meta-análisis no debe faltar la valoración de la calidad metodológica de los estudios primarios con objeto de comprobar si ésta puede estar estadísticamente relacionada con los tamaños del efecto. Si existe una relación estadística entre calidad del estudio y tamaño del efecto, entonces ésta es evidencia de que los estudios con una pobre calidad pueden estar ofreciendo estimaciones sesgadas de los efectos paramétricos. Es por ello que la evaluación de la calidad de los estudios primarios constituye un elemento crítico en un meta-análisis.

Se han propuesto en la literatura más de 200 escalas para evaluar la calidad metodológica de los estudios primarios (Conn y Rantz, 2003; Deeks, Dinnes, D'Amico *et al*., 2003; Saunders, Soomro, Buckingham *et al*., 2003). La mayoría de ellas proceden del ámbito médico y, en consecuencia, su aplicabilidad al contexto de las ciencias sociales puede ser limitado. De hecho, muchas de estas escalas están diseñadas para valorar la calidad de los ensayos clínicos aleatorizados (ECAs), mientras que en ciencias sociales la asignación aleatoria de las unidades experimentales a las condiciones experimentales no está tan generalizada debido a distintas razones (e.g., éticas, logísticas, etc.).

El objetivo de esta investigación es demostrar, mediante una aplicación a un meta-análisis real, cómo determinados aspectos relacionados con la calidad metodológica de los estudios primarios pueden estar estadísticamente relacionados con los tamaños del efecto obtenidos en dichos estudios. Para ello, hemos aplicado una escala de calidad elaborada por nuestro equipo de investigación a un conjunto de estudios de un meta-análisis real. Nuestros resultados también nos permitirán comprobar si la escala de calidad es sensible para detectar diferencias en la calidad metodológica y cómo éstas afectan a las estimaciones de los efectos.

Nuestra escala de calidad está compuesta por diez ítems dicotómicos, cada uno de los cuales hace referencia a algún aspecto relativo a la calidad metodológica del estudio primario y, fundamentalmente, relacionada con la validez interna del diseño. Cada ítem es valorado como 0-1 según que el estudio primario cumpla (1) o no cumpla (0) con el ítem: (1) asignación aleatoria de los participantes a las condiciones experimentales; (2) igualación de los grupos en el pretest en las variables dependientes y en las variables de control (al menos en el 90% de ellas); (3) inclusión de medidas pretest; (4) inclusión de un grupo de control placebo psicológico; (5) inclusión de un grupo de control placebo farmacológico; (6) enmascaramiento del evaluador; (7) uso de instrumentos de medida debidamente validados; (8) tamaño muestral total en el postest por encima de la media del conjunto de estudios; (9) mortalidad diferencial entre los grupos tratado y de control igual o inferior al 10%, y (10) uso de análisis por intención de tratar.

Nuestra hipótesis es que cuando un estudio primario no cumple con los criterios de calidad definidos en cada uno de estos ítems, la estimación del tamaño del efecto ofrecida por dicho estudio exhibirá un sesgo positivo del verdadero efecto en la población y, en consecuencia, la puntuación total de la escala de calidad estará negativamente relacionada con el tamaño del efecto.

MÉTODO

Muestra

Para alcanzar nuestro objetivo aplicamos el método de caso, mediante el cual elegimos un meta-análisis sobre la eficacia de los tratamientos psicológicos del trastorno de pánico con o sin agorafobia realizado por nuestro equipo (Sánchez-Meca,

Rosa-Alcázar, Marín-Martínez y Gómez-Conesa, 2010). Este meta-análisis incluyó 65 estudios evaluativos que comparaban un grupo tratado con un grupo de control.

Procedimiento

Con cada uno de los estudios se aplicó la escala de calidad arriba descrita y, para comprobar la relación entre los ítems de calidad y el tamaño del efecto, se calculó la diferencia media tipificada con las medias y desviaciones típicas de los dos grupos (tratado y control) obtenidos en el postest sobre la variable dependiente más importante del estudio, que fue la medida con escalas que evaluaban los síntomas de pánico y/o evitación agorafóbica. Además, dado que el ítem 9 (mortalidad diferencial) estaba dicotomizado, añadimos a nuestros análisis la mortalidad global del estudio en el postest (tomando conjuntamente los dos grupos) y la mortalidad diferencial entre los dos grupos. Finalmente, también se calculó la diferencia media tipificada entre los dos grupos tomando las medidas del pretest, con objeto de comprobar si este índice estaba estadísticamente relacionado con el tamaño del efecto alcanzado en el postest. Este índice está directamente relacionado con l ítem 2, igualación de los grupos en el pretest, aunque sólo hacía referencia a la variable dependiente fundamental.

Análisis estadístico

Para comprobar la relación estadística de cada ítem dicotómico de la escala de calidad con el tamaño del efecto, se aplicó un ANOVA de efectos mixtos, según el cual cada tamaño del efecto se ponderó por la inversa de su varianza. Para comprobar la relación estadística de la puntuación total de la escala de calidad (suma de los diez ítems), de la mortalidad global, de la mortalidad diferencial y de la diferencia media tipificada en el pretest con el tamaño del efecto, se aplicaron meta-regresiones simples de efectos mixtos. Finalmente, se aplicó una meta-regresión múltiple de efectos mixtos incluyendo todos los ítems de calidad, la mortalidad global, la mortalidad diferencial y la diferencia media tipificada en el pretest, para comprobar la potencia explicativa de todos ellos sobre los tamaños del efecto y detectar cuál/es de ellos son los que mayor relación estadística presentan (Borenstein, Hedges, Higgins y Rothstein, 2009).

Resultados

De los 10 ítems de que consta la escala hubo dos que no pudieron analizarse debido a que todos los estudios puntuaron positivamente en ellos: los ítems 3 (inclusión de medidas pretest) y 7 (uso de instrumentos de medida debidamente validados). La Tabla 1 presenta los resultados de los ANOVAs ponderados aplicados con cada uno de los ocho ítems restantes de la escala de calidad, tomando como variable dependiente el tamaño del efecto en el postest. De los ocho ítems analiza-

dos, dos de ellos alcanzaron un resultado estadísticamente significativo ($p < .05$) y otros dos obtuvieron un resultado marginalmente significativo ($p < .10$). Los dos ítems significativos fueron el enmascaramiento del evaluador ($p = .007$) y el uso de un grupo de control placebo farmacológico ($p = .011$). Los dos ítems con resultados marginalmente significativos fueron la asignación aleatoria ($p = .062$) y la mortalidad diferencial ($p = .072$). De estos cuatro marcadores de calidad, sólo uno de ellos obtuvo un resultado en la línea de nuestra hipótesis de partida: un menor tamaño del efecto cuando se cumplió el ítem de calidad. Éste fue el caso del uso de un grupo de control placebo farmacológico, que presentó un efecto medio más bajo cuando el tratamiento se comparó con un placebo farmacológico que cuando se comparó con un control en lista de espera. Sin embargo, los otros tres marcadores exhibieron un resultado contrario a nuestra hipótesis: un mayor tamaño del efecto cuando se cumplió el ítem. En todos los casos, el porcentaje de varianza inter-estudios explicada por los ítems fue muy bajo (por debajo del 10%), lo que indica una escasa contribución de estos ítems en la explicación de la variabilidad de los tamaños del efecto.

Tabla 1. Resultados de los ANOVAs aplicados con cada ítem de la escala de calidad

Item de calidad	k	Media (I.C. 95%)	Resultados del ANOVA
1. Asignación aleatoria:			$Q_B(1) = 3{,}47$; $p = .062$
No	8	0,587 (0,057; 1,117)	$Q_W(63) = 301{,}96$; $p < .001$
Sí	57	1,126 (0,923; 1,328)	$R^2 = 0{,}011$
2. Igualación pretest:			$Q_B(1) = 1{,}54$; $p = .214$
No	9	0,744 (0,215; 1,274)	$Q_W(63) = 303{,}16$; $p < .001$
Sí	56	1,104 (0,901; 1,307)	$R^2 = 0{,}0$
4. Control placebo psicológico:			$Q_B(1) = 0{,}38$; $p = .535$
No	57	1,081 (0,877; 1,285)	$Q_W(63) = 310{,}76$; $p < .001$
Sí	8	0,897 (0,354; 1,441)	$R^2 = 0{,}0$
5. Control placebo farmacológico:			$Q_B(1) = 6{,}52$; $p = .011$
No	52	1,182 (0,973; 1,392)	$Q_W(63) = 289{,}72$; $p < .001$
Sí	13	0,599 (0,204; 0,995)	$R^2 = 0{,}062$
6. Enmascaramiento del evaluador:			$Q_B(1) = 7{,}19$; $p = .007$
No	43	0,874 (0,644; 1,105)	$Q_W(63) = 299{,}76$; $p < .001$
Sí	22	1,419 (1,094; 1,744)	$R^2 = 0{,}024$
8. Tamaño muestral:			$Q_B(1) = 0{,}08$; $p = .779$
< 34	32	1,089 (0,803; 1,375)	$Q_W(63) = 310{,}96$; $p < .001$
≥ 34	33	1,034 (0,778; 1,291)	$R^2 = 0{,}0$
9. Mortalidad :			$Q_B(1) = 3{,}23$; $p = .072$
> 10%	38	0,920 (0,678; 1,163)	$Q_W(63) = 306{,}93$; $p < .001$
≤ 10%	27	1,278 (0,972; 1,584)	$R^2 = 0{,}0$
10. Análisis por intención de tratar:			$Q_B(1) = 0{,}39$; $p = .533$
No	40	1,105 (0,864; 1,346)	$Q_W(63) = 303{,}36$; $p < .001$
Sí	25	0,982 (0,677; 1,286)	$R^2 = 0{,}007$

La Tabla 2 presenta los resultados de las meta-regresiones simples ponderadas de cada predictor metodológico continuo con el tamaño del efecto. Se obtuvo una relación estadísticamente significativa con los predictores '*d* en el pretest' (p = .005) y mortalidad global (p = .0002), exhibiendo en ambos casos una relación negativa con el tamaño del efecto. Esta relación negativa también va contra nuestra hipótesis de partida de un menor tamaño del efecto cuanto mejor es la calidad de los estudios. Cabe destacar el porcentaje de vaianza explicada por la mortalidad global en el postest, con un 11,1%.

Tabla 2. Resultados de las meta-regresiones simples aplicadas con cada predictor

Predictor	k	b_j	ET_j	Z	p	R^2
***d* en el pretest**	61	-0,802	0,285	-2,81	.005	0,043
Mortalidad global	65	-2,645	0,714	-3,70	.0002	0,111
Mortalidad diferencial	58	-1,968	0,864	-1,38	.166	0,009
Total Escala	65	0,114	0,083	1,37	.171	0,0

Finalizamos los análisis estadísticos aplicando un modelo de meta-regresión múltiple de efectos mixtos, tomando como predictores los ocho ítems dicotómicos de nuestra escala de calidad más el índice *d* en el pretest y la mortalidad global, ya que estas dos últimas variables presentaron una relación estadísticamente significativa con el tamaño del efecto en sus meta-regresiones simples respectivas. La variable dependiente continuó siendo el tamaño del efecto en el postest. El modelo completo presentó una relación estadísticamente significativa [$Q_R(10) = 41{,}715$, $p < .0001$; $R^2 = 0{,}264$], aunque continuó estando mal especificado [$Q_E(50) = 196{,}416$, $p < .0001$]. El análisis de los predictores del modelo reveló una relación estadísticamente significativa para el índice *d* en el pretest ($b_j = -0{,}881$; $p = .0021$), la mortalidad global en el postest ($b_j = -3{,}274$; $p = .0026$) y el ítem 6, 'enmascaramiento del evaluador' ($b_j = 0{,}701$; $p = .0008$). Estos tres predictores presentaron coeficientes de regresión parcializados con signo contrario al esperado desde nuestra hipótesis de un mayor tamaño del efecto cuando no se cumple el criterio de calidad.

CONCLUSIONES

El propósito de nuestro estudio fue analizar, mediante la aplicación del método de caso, las relaciones existentes entre marcadores de calidad metodológica de los estudios primarios de un meta-análisis y el tamaño del efecto obtenido en dichos estudios. Para ello, aplicamos nuestra escala de calidad metodológica de 10 ítems dicotómicos a los 66 estudios empíricos de un meta-análisis sobre la eficacia de los tratamientos psicológicos del trastorno de pánico con o sin agorafobia.

Contrario a nuestra hipótesis original, basada en los resultados de investigaciones previas de otros autores, en este meta-análisis apenas encontramos alguna relación estadística entre calidad metodológica y tamaño del efecto en la línea de nuestra hipótesis. Antes bien, obtuvimos relación estadística contraria con algunos

ítems de la escala de calidad, relación estadística favorable a nuestra hipótesis con unos pocos ítems y ausencia de relación estadística con la mayoría de ellos.

Lo que estos resultados ponen de manifiesto es que no es posible generalizar a cualquier ámbito de estudio la existencia de sobreestimaciones de los efectos cuando los estudios primarios tienen baja calidad metodológica. La relación entre calidad metodológica y tamaño del efecto es una cuestión empírica que varía de una base de estudios a otra y, en consecuencia, sólo puede dilucidarse analizando su relación en cada caso concreto. Este resultado refuerza más aún la gran importancia que tiene valorar la calidad metodológica de los estudios primarios cuando se hace un meta-análisis con objeto de poder detectar infraestimaciones o sobreestimaciones de los efectos dependiendo de diferentes marcadores de calidad. Este resultado también nos lleva a desaconsejar la práctica, habitual en algunos ámbitos de las ciencias de la salud, de desechar estudios primarios con baja calidad metodológica sin ni siquiera comprobar empíricamente si dichos estudios están ofreciendo estimaciones sesgadas de los efectos.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación (Proyecto nº PSI2009/12172).

Referencias

Borenstein, M., Hedges, L.V., Higgins, J.P.T. y Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Conn, V. S. y Rantz, M.J. (2003). Research methods: Managing primary study quality in meta-analyses. *Research in Nursing and Health, 26*, 322-333.

Deeks, J.J., Dinnes, J., D'Amico, R., Sowden, A.J., Sakarovitch, C., Song, F., Petticrew, M. y Altman, D.G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27).

Saunders, L., Soomro, G., Buckingham, J., Jamtvedt, G. y Raina, P. (2003). Assessing the methodological quality of nonrandomized intervention studies. *Western Journal of Nursing Research, 25*, 223-237.

EL ENFOQUE META-ANALÍTICO DE GENERALIZACIÓN DE FIABILIDAD

Coordinador: José Antonio López Pina
Universidad de Murcia

Una de las más recientes extensiones del meta-análisis ha sido el desarrollo del enfoque de generalización de la fiabilidad (GF), por el que es posible integrar estadísticamente las estimaciones de la fiabilidad obtenidas en repetidas aplicaciones de un mismo test psicométrico. Dado que desde la teoría clásica de tests la fiabilidad no es una propiedad inmutable sino que varía en función de la composición y variabilidad de la muestra de participantes sobre la que se aplica, el meta-análisis constituye una herramienta ideal para estudiar cómo varían las estimaciones de la fiabilidad de un test en sus diversas aplicaciones a muestras concretas de participantes. Un estudio GF es, pues, un meta-análisis en el que se reúnen los estudios que han aplicado un determinado test y que reportan alguna estimación de la fiabilidad con los propios datos de la muestra. Con el conjunto de coeficientes de fiabilidad es posible a continuación estimar la fiabilidad media de todas las aplicaciones y examinar qué características de las muestras, del test y del contexto de aplicación del mismo pueden estar afectando a la variabilidad de dichas estimaciones. Desde que Vacha-Haase (1998) propusiera hace 13 años el enfoque meta-analítico de generalización de la fiabilidad, se han propuesto en la literatura mejoras metodológicas para optimizar la realización de estudios GF. Sin embargo, quedan todavía múltiples lagunas metodológicas por resolver y, además, no existe en la actualidad un consenso sobre cuál es el mejor método para analizar estadísticamente los datos procedentes de un estudio GF. El propósito de este simposio es presentar algunos avances metodológicos y estadísticos en el ámbito de la generalización de la fiabilidad. En la primera comunicación José A. López López presenta los resultados de un estudio de simulación Monte Carlo dirigido a comparar el funcionamiento, en términos del control de la tasa de error Tipo I, de diferentes métodos estadísticos propuestos recientemente en la literatura meta-analítica para comprobar la significación estadística de variables moderadoras sobre los coeficientes alfa asumiendo un modelo de meta-regresión de efectos mixtos. En la segunda comunicación, José A. López Pina presenta los resultados de una aplicación a los datos reales de un estudio GF de diez modelos estadísticos diferentes propuestos en la literatura para estimar la fiabilidad media y examinar el influjo de variables moderadoras. En la tercera comunicación Juan Botella presenta varios métodos recientemente propuestos para meta-analizar curvas ROC en el ámbito de la investigación experimental y los aplica a un ejemplo real concreto. Aunque este trabajo no se enmarca-

ría en sentido estricto dentro del enfoque GF, bien es cierto que el meta-análisis de curvas ROC se está utilizando también en el estudio de la precisión (es decir, la fiabilidad) de pruebas diagnósticas, por lo que esta aportación puede perfectamente enmarcarse dentro del enfoque GF. Por último, Julio Sánchez Meca presenta un método para incrementar el número de estudios meta-analizables en un estudio GF mediante la aplicación de la fórmula KR-21 para estimar el coeficiente alfa cuando los estudios no reportan ninguna estimación de la fiabilidad. Dicho método es aplicado a una base de datos real procedente de un estudio GF.

PALABRAS CLAVE: Generalización de la fiabilidad, Meta-análisis, Coeficiente alfa, Modelos estadísticos, Curvas ROC.

ALTERNATIVAS METODOLÓGICAS PARA EL AJUSTE DE MODELOS MIXTOS DE META-REGRESIÓN DENTRO DEL ENFOQUE DE GENERALIZACIÓN DE LA FIABILIDAD

José Antonio López-López[1], Juan Botella[2], Julio Sánchez-Meca[1] y Fulgencio Marín-Martínez[1]

[1] Universidad de Murcia

[2] Universidad Autónoma de Madrid

Correo electrónico: josealopezlopez@um.es

Resumen

Uno de los objetivos del enfoque de generalización de la fiabilidad consiste en identificar predictores de la variabilidad entre los coeficientes integrados. Este estudio comparó mediante simulación Monte Carlo el funcionamiento de diferentes alternativas metodológicas para el análisis de meta-regresión de efectos mixtos. En concreto, se combinaron cuatro transformaciones de los coeficientes alfa, dos estimadores de la varianza inter-estudios residual y dos métodos para estimar la matriz de covarianzas de los coeficientes de regresión. Los criterios comparativos presentados aquí son la tasa de error Tipo I y potencia estadística. Los resultados mostraron principalmente un mejor funcionamiento en ambos casos utilizando el método de Knapp y Hartung (2003) para estimar las varianzas de los coeficientes de regresión, mientras que el método para transformar los coeficientes y el estimador de la varianza inter-estudios residual no mostraron una gran influencia en la conclusión estadística.

La fiabilidad es una de las propiedades psicométricas más importantes a la hora de elegir un cuestionario para su administración en un contexto determinado. Sin embargo, no es una propiedad inmutable del test, sino que varía a lo largo de diferentes aplicaciones del mismo (e.g., Crocker y Algina, 1986). El enfoque meta-analítico de generalización de la fiabilidad (GF; Vacha-Haase, 1998) tiene como principal objetivo la integración de un conjunto de estimaciones de la fiabilidad obtenidas para un mismo test, obteniendo un valor promedio que puede considerarse más o menos representativo del test en cuestión, en función del modelo estadístico adoptado (Sánchez-Meca, López-Pina y López-López, 2009). Además, dado que habitualmente los coeficientes de fiabilidad presentan más variabilidad de la que sería esperable por mero error de muestreo, otro objetivo de los estudios GF consiste en identificar características de las muestras y de los estudios capaces de explicar parte de esta heterogeneidad.

Existen varios aspectos metodológicos para los cuales aún no se ha alcanzado un consenso dentro del enfoque GF. Uno de ellos tiene que ver con la transforma-

ción de los coeficientes. Mientras que algunos autores aconsejan analizar los coeficientes brutos (e.g., Henson y Thompson, 2002; Vacha-Haase, 1998), varios procedimientos de transformación han sido propuestos para normalizar su distribución y estabilizar sus varianzas (Bonett, 2002; Hakstian y Whalen, 1976). Otro aspecto para el que se han empleado diversas soluciones dentro del enfoque es la ponderación de los coeficientes. En algunos estudios GF previos se analizaron los coeficientes sin ponderar (e.g., Vacha-Haase, 1998). Cuando se ha utilizado algún esquema de ponderación, a veces se han ponderado los coeficientes por sus tamaños muestrales (e.g., Yin y Fan, 2000), mientras que en otras ocasiones el factor de ponderación ha sido la inversa de la varianza (e.g. Sánchez-Meca, López-Pina, López-López, Marín-Martínez, Rosa-Alcázar y Gómez-Conesa, 2011). Al llevar a cabo un análisis de moderadores ponderando por la inversa de la varianza, las dos alternativas estadísticas más frecuentes consisten en asumir un modelo de efectos fijos o de efectos mixtos. Hoy en día se considera que el modelo de efectos mixtos es una opción más realista, además de permitirnos generalizar los resultados más allá de las muestras seleccionadas para el meta-análisis (e.g., Raudenbush, 2009).

El modelo de efectos mixtos asume que los coeficientes integrados en el meta-análisis constituyen una muestra aleatoria de una superpoblación de coeficientes paramétricos. En la práctica, esto implica dos componentes de varianza: la varianza intra-estudio (v_i), o error muestral para cada coeficiente estimado, y la varianza inter-estudios residual (τ^2), resultante del muestreo de coeficientes a partir de una superpoblación. Ambos componentes deben ser estimados, aunque una práctica habitual en meta-análisis consiste en asumir que el valor de v_i es conocido. Por el contrario, el valor de τ^2 es siempre una estimación, y existen diferentes procedimientos para calcular la misma.

Por otra parte, poner a prueba la relación de una o más variables moderadoras con los coeficientes de fiabilidad exige el contraste de las respectivas pendientes. El método tradicionalmente empleado para este fin ha sido criticado (e.g., Sidik y Jonkman, 2005), ya que su funcionamiento está fuertemente condicionado por la precisión con que se hayan estimado las varianzas muestrales. Para paliar esta debilidad, Knapp y Hartung (2003) propusieron un nuevo método añadiendo un factor de corrección al método tradicional. En su estudio de simulación, utilizando *odds ratios* como índice del tamaño del efecto, la nueva propuesta mejoró los resultados del método tradicional en términos de ajuste al nivel nominal de significación. No obstante, el método de Knapp-Hartung no ha sido empleado hasta la fecha en ningún estudio GF y, puesto que en estos últimos se trabaja con coeficientes de fiabilidad como variable dependiente (en lugar de índices de riesgo como las *odds ratios*), los resultados podrían diferir.

Este trabajo comparó, mediante simulación Monte Carlo, diferentes alternativas metodológicas para el ajuste de modelos de meta-regresión de efectos mixtos. Específicamente, se combinaron dos estimadores de la varianza residual, las extensiones de los métodos de DerSimonian y Laird (DL) y de máxima verosimilitud restringida (REML), cuyas fórmulas pueden encontrarse en otros textos (e.g.,

Raudenbush, 2009); dos métodos para estimar las varianzas de los coeficientes de regresión, concretamente el método tradicional frente a método de Knapp-Hartung, cuyo cálculo se detalla en otros documentos (e.g., Knapp y Hartung, 2003; Sidik y Jonkman, 2005); y cuatro transformaciones de los coeficientes de fiabilidad: coeficientes brutos, Z de Fisher y transformaciones de Hakstian y Whalen (1976) y Bonett (2002). Este estudio se centró en coeficientes alfa, que son los que más frecuentemente reportan los estudios primarios y, por ello, constituyen la principal variable dependiente en la mayoría de los estudios GF publicados hasta la fecha.

MÉTODO

La simulación se llevó a cabo de manera similar a la descrita en Botella y Suero (en prensa). Los tamaños muestrales de cada estudio, N_i, se generaron a partir de una distribución log-normal con media 150. La asimetría de esta distribución se manipuló con valores 1, 2 y 3, de acuerdo con las bases de datos de estudios GF previos (e.g., López-Pina, Sánchez-Meca y Rosa-Alcázar, 2009; Sánchez-Meca *et al.*, 2011). También se manipuló el número de estudios de cada meta-análisis, $k = \{15, 30, 60\}$. Además, para la pendiente paramétrica, se consideraron dos escenarios: en el primero, se generó un predictor independiente con distribución N(0, 1), de manera que el valor esperado para la pendiente era 0; en el segundo caso, el componente de error en las puntuaciones del test se calculó como una función del predictor, resultando en una pendiente paramétrica media de 0,01348 (rango 0,01346-0,01350).

En una primera fase, se generaron las puntuaciones verdaderas de la población de cada estudio para un test de 20 ítems, a partir de una distribución normal multivariada con 10000 casos. Seguidamente, se generaron las puntuaciones de error para cada ítem, sumando ambas para obtener las puntuaciones verdaderas. Los coeficientes alfa paramétricos se calcularon a partir de esta matriz de 10000 sujetos, mientras que los valores para cada estudio del meta-análisis se calcularon extrayendo muestras de N_i sujetos de la matriz completa. Se generaron 10000 meta-análisis para cada condición.

Las variables de resultado que se mostrarán aquí son la proporción de rechazos de la hipótesis nula cuando la pendiente paramétrica valía 0 (error Tipo I) y cuando su valor era mayor que 0 (potencia estadística). Además, se evaluó el sesgo y la eficiencia de las estimaciones de la pendiente paramétrica con los diferentes métodos, aunque los resultados mostraron un funcionamiento apropiado y similar en todos los casos, por lo que no serán recogidos aquí.

RESULTADOS

Error Tipo I

La Tabla 1 muestra los resultados con el estimador DL. En este documento no se presentan datos obtenidos con REML, puesto que los valores fueron muy simi-

lares y mostraron las mismas tendencias observadas con el estimador DL. Con un nivel de confianza del 95%, se observó una mayor proximidad de las tasas de error Tipo I al nivel nominal (5%) al utilizar el método de Knapp-Hartung, en comparación con el método tradicional. Cuando se aplicó el método de Knapp-Hartung, los resultados estuvieron próximos a la tasa nominal de rechazo independientemente del resto de factores metodológicos y condiciones manipuladas.

Tabla 1. Tasa de error Tipo I utilizando el estimador DL

k		**15**			**30**			**60**		
Asimetría		1	2	3	1	2	3	1	2	3
Coeficiente alfa bruto	TRA	,017	,020	,021	,014	,018	,020	,017	,018	,021
	KH	,048	,053	,054	,049	,050	,050	,050	,053	,055
Z-Fisher	TRA	,006	,007	,007	,006	,006	,006	,005	,005	,006
	KH	,046	,049	,049	,050	,048	,043	,048	,048	,048
Hakstian-Whalen	TRA	,017	,020	,023	,017	,019	,019	,019	,020	,021
	KH	,046	,049	,049	,049	,049	,044	,047	,049	,048
Bonett	TRA	,016	,020	,022	,018	,020	,019	,020	,020	,021
	KH	,046	,048	,049	,049	,047	,043	,048	,048	,047

k: número de estudios. *Asimetría*: grado de asimetría en la distribución de tamaños muestrales. TRA y KH: método tradicional y corrección de Knapp-Hartung, respectivamente, para estimar la matriz de covarianzas de los coeficientes de regresión.

Potencia estadística

Los resultados de la Tabla 2 mostraron una clara influencia directa del número de estudios con la potencia estadística, mientras que la asimetría en la distribución de tamaños muestrales mostró una ligera relación inversa. En cuanto a los métodos comparados, el método de Knapp-Hartung obtuvo valores de potencia sistemáticamente más altos que el método tradicional.

Tabla 2. Tasa de potencia estadística utilizando el estimador DL

k		**15**			**30**			**60**		
Asimetría		1	2	3	1	2	3	1	2	3
Coeficiente alfa bruto	TRA	,266	,271	,259	,561	,552	,556	,884	,871	,875
	KH	,369	,363	,347	,682	,666	,658	,942	,929	,925
Z-Fisher	TRA	,171	,170	,166	,425	,418	,422	,810	,795	,799
	KH	,368	,364	,348	,684	,675	,667	,942	,931	,930
Hakstian-Whalen	TRA	,283	,282	,271	,587	,577	,580	,902	,890	,892
	KH	,370	,363	,341	,685	,671	,661	,943	,932	,929
Bonett	TRA	,284	,281	,270	,586	,578	,581	,901	,892	,894
	KH	,369	,362	,345	,684	,674	,664	,942	,931	,930

k: número de estudios. *Asimetría*: grado de asimetría en la distribución de tamaños muestrales. TRA y KH: método tradicional y corrección de Knapp-Hartung, respectivamente, para estimar la matriz de covarianzas de los coeficientes de regresión.

DISCUSIÓN

Este trabajo se centró en el análisis de moderadores cuantitativos a partir de modelos mixtos de meta-regresión simple en el enfoque GF, empleando coeficientes alfa como variable dependiente. Diferentes alternativas para llevar a cabo este tipo de análisis fueron comparadas mediante simulación Monte Carlo, siendo los criterios comparativos la tasa empírica de error Tipo I y la potencia estadística para cada método al contrastar la significación de la pendiente del modelo.

En cuanto a las condiciones consideradas en esta simulación, el número de estudios, *k*, mostró una clara relación directa con la precisión de los resultados, algo habitual en meta-análisis (e.g., Botella y Gambara, 2002). Por otra parte, la asimetría en la distribución de los tamaños muestrales mostró una leve relación inversa con la potencia estadística. Esto último constituye un hallazgo novedoso aunque, dado que la asimetría en la distribución de tamaños muestrales no ha sido manipulada en los otros estudios de simulación del enfoque GF realizados hasta la fecha, se requieren simulaciones más exhaustivas en este factor para poder plantear recomendaciones fundamentadas.

En cuanto a los métodos comparados en este trabajo, la transformación de los coeficientes y el estimador de la varianza inter-estudios residual mostraron una escasa influencia en los criterios comparativos del estudio. Dada la distribución marcadamente asimétrica del coeficiente alfa, nuestra recomendación es aplicar alguna transformación previa a los análisis. Frente a la transformación Z de Fisher, que ha sido hasta ahora la más empleada en los estudios GF (cf. Sánchez-Meca, López-Pina y López-López, 2008), aconsejamos las transformaciones de Hakstian-Whalen y de Bonett, que han sido específicamente diseñadas para coeficientes alfa.

Por otra parte, el método para estimar la matriz de covarianzas de los coeficientes de regresión mostró una influencia clara en los resultados. En concreto, tanto para error Tipo I como para potencia estadística, las tasas obtenidas mostraron un mejor funcionamiento de la corrección propuesta por Knapp y Hartung (2003) frente al método tradicional. Según nuestros datos, esta corrección aún no ha sido aplicada en ningún estudio GF, por lo que los resultados obtenidos aquí deben suponer un incentivo para que los investigadores de este campo se planteen su empleo.

El enfoque GF es, como ya se ha enfatizado, una aplicación reciente del meta-análisis que requiere aún de numerosas contribuciones metodológicas para alcanzar un mayor refinamiento. En esa línea, los resultados de este trabajo apoyan claramente el uso del ajuste de Knapp-Hartung para contrastar la significación de los predictores en modelos de meta-regresión de efectos mixtos.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por la Fundación Séneca de la C.A. de la Región de Murcia (Proyecto nº 08650/PHCS/08).

REFERENCIAS

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*, 335-340.

Botella, J. y Gambara, H. (2002). *¿Qué es el meta-análisis?* Madrid: Biblioteca Nueva.

Botella, J., y Suero, M. (en prensa). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology*.

Crocker, L., y Algina, J. (1986). *Introduction to classical and modern test theory.* Nueva York: Holt, Rinehart and Winston.

Hakstian, A. R., y Whalen, T. E. (1976). A *k*-sample significance test for independent alfa coefficients. *Psychometrika, 41,* 219–231.

Henson, R. K., y Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting «reliability generalization» studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

Knapp, G., y Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics In Medicine, 22*, 2693-2710.

López-Pina, J. A., Sánchez-Meca, J., y Rosa-Alcázar, A. I. (2009). The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology, 9*, 143-159.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. En H. Cooper, L. V. Hedges, y J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2ª ed.) (pp. 295-315). Nueva York: Russell Sage Foundation.

Sánchez-Meca, J., López-Pina, J. A., y López-López, J. A. (2008). Una revisión de los estudios meta-analíticos de generalización de la fiabilidad. *Escritos de Psicología,1-2*, 107-118.

Sánchez-Meca, J., López-Pina, J. A., y López-López, J. A. (2009). Generalización de la fiabilidad: Un enfoque meta-analítico aplicado a la fiabilidad. *Fisioterapia, 31*, 262-270.

Sánchez-Meca, J., López-Pina, J. A., López-López, J. A., Marín-Martínez, F., Rosa-Alcázar, A. I., y Gómez-Conesa, A. (2011). The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology*, *11*, 473-493.

Sidik, K., y Jonkman, J. N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, *15*, 823-838.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58,* 6-20.

Yin, P., y Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-223.

MODELOS ESTADÍSTICOS EN LOS ESTUDIOS META-ANALÍTICOS DE GENERALIZACIÓN DE LA FIABILIDAD

José A. López-Pina[1], Julio Sánchez-Meca[1], José A. López-López[1], Fulgencio Marín-Martínez[1], Rosa Mª Núñez-Núñez[2], Ana I. Rosa-Alcázar[1] y Antonia Gómez-Conesa[1]

[1] Universidad de Murcia

[2] Universidad Miguel Hernández de Elche

Correo electrónico: jlpina@um.es

Resumen

Los precursores del enfoque de generalización de la fiabilidad (GF) no han llegado a un acuerdo con respecto al método analítico más adecuado para integrar un conjunto de coeficientes de fiabilidad calculados en diferentes aplicaciones de un test psicométrico. En este trabajo se presenta una comparación de seis modelos estadísticos para promediar coeficientes alfa y para analizar la influencia de variables moderadoras. Los diez modelos en cuestión son: (a) la aplicación de métodos estadísticos convencionales sobre los coeficientes alfa; (b) el modelo de efecto fijo aplicado sobre los coeficientes alfa transformados mediante la transformación de Bonett; (c) el modelo de efectos aleatorios aplicado sobre esta misma transformación de los coeficientes alfa; (d) el modelo de efectos aleatorios mejorado según la propuesta de Knapp y Hartung; (e) el modelo de Hunter y Schmidt basado en la ponderación por el tamaño muestral de cada estudio, y (f) el modelo de coeficientes variables aconsejado por Bonett. Para alcanzar nuestro objetivo aplicamos los seis modelos estadísticos a un estudio de GF realizado sobre la *Escala de Obsesiones y Compulsiones de Yale-Brown* (Y-BOCS). Los resultados obtenidos con los diferentes métodos muestran importantes discrepancias que pueden afectar a las conclusiones.

Los estudios de generalización de la fiabilidad (GF) son un tipo de meta-análisis psicométrico que tienen por objetivo integrar estadísticamente los coeficientes de fiabilidad, definidos desde la teoría clásica de los test, obtenidos en diferentes administraciones de un mismo test a diferentes muestras procedentes de diferentes poblaciones y en distintos contextos (Thompson, 2003; Vacha-Haase, 1998). Este tipo de meta-análisis permite estimar la fiabilidad media de las puntuaciones del test y explorar qué características de los estudios (de los participantes y del contexto de aplicación) pueden dar cuenta de la variabilidad de los coeficientes (Botella, Suero y Gambara, 2010; Sánchez-Meca, López-Pina y López-López, 2008, 2009).

Sobre la base de datos meta-analítica, se aplican técnicas estadísticas para obtener un coeficiente de fiabilidad medio con su intervalo de confianza, estadísticos que cuantifican el grado de heterogeneidad y pruebas estadísticas dirigidas a comprobar qué características de los estudios están estadísticamente relacionadas con los coeficientes de fiabilidad. Se han propuesto en la literatura meta-analítica diferentes modelos estadísticos para integrar un conjunto de coeficientes alfa, pero los precursores del enfoque de GF no han alcanzado un acuerdo sobre qué modelo(s) se debe(n) utilizar. De hecho, aconsejan que el enfoque de GF no sea considerado como un método monolítico, sino que se deja libertad a los investigadores para que adopten los métodos estadísticos que consideren oportunos (Henson y Thompson, 2002). Esta práctica, sin embargo, hace difícil la comparabilidad de los resultados de diferentes estudios de GF y da la impresión de que 'todo vale' a la hora de analizar los datos. Además, es habitual en los estudios de GF no informar del modelo estadístico que están asumiendo para realizar los cálculos estadísticos. Si diferentes modelos estadísticos dan lugar a resultados discrepantes, entonces la elección de éste cobra una relevancia fundamental para interpretar adecuadamente dichos resultados.

El propósito de este estudio es, mediante el método de caso, comprobar en qué grado los resultados de diferentes modelos estadísticos aplicados sobre una misma base de coeficientes de fiabilidad dan lugar a resultados discrepantes. Dado que el coeficiente alfa es el más utilizado en los estudios primarios para estimar la fiabilidad de las puntuaciones de los tests, nuestra investigación se centra en dicho coeficiente, si bien buena parte de las conclusiones son generalizables a otros coeficientes de fiabilidad.

MODELOS ESTADÍSTICOS EN GF

Los tres modelos estadísticos propuestos en la literatura sobre GF son el modelo de efecto fijo, el de coeficientes variables y el de efectos aleatorios (Bonett, 2010; Borenstein, Hedges, Higgins y Rothstein, 2010). En todos los casos, se asume que disponemos de un conjunto de *k* coeficientes de fiabilidad independientes (transformados o no), que representaremos por T_i, obtenidos en una serie de estudios empíricos. El coeficiente de fiabilidad paramétrico que estima cada coeficiente de fiabilidad muestral queda representado por α_i.

El *modelo de efecto fijo* asume que todos los coeficientes de fiabilidad, T_i, obtenidos en los estudios están estimando a un coeficiente alfa paramétrico común a todos ellos, α, de forma que el modelo matemático puede formularse como $T_i = \alpha + u_i$, siendo u_i los errores de muestreo aleatorio debidos al hecho de que cada estudio ha calculado el coeficiente de fiabilidad con una muestra (aleatoria) diferente. Desde este modelo se asume que sólo nos interesa generalizar los resultados a una población de estudios con características idénticas a las de los estudios empíricos meta-analizados.

Se han aplicado en algunos estudios de GF dos modalidades dentro de este modelo para calcular el coeficiente de fiabilidad medio, su intervalo de confianza y el análisis de variables capaces de explicar la variabilidad de los coeficientes. Uno de ellos fue propuesto por Vacha-Haase (1998) en su artículo seminal sobre GF y consiste en

no transformar los coeficientes alfa y no ponderarlos en los análisis estadísticos, mientras que el otro se basa en la propuesta de Hedges y Olkin (1985), consistente en transformar los coeficientes de fiabilidad mediante la propuesta de Bonett (2002) y en ponderar los coeficientes en función de la inversa de su varianza intra-estudio.

El *modelo de coeficientes variables* ha sido recientemente propuesto por Bonett (2010) y asume que cada coeficiente de fiabilidad T_i estima a un coeficiente paramétrico diferente, α_i, por lo que el modelo estadístico se formula como $T_i = \alpha_i + u_i$. Este modelo propone no transformar los coeficientes de fiabilidad para el cálculo del coeficiente medio, pero sí para obtener su intervalo de confianza y para el análisis de variables moderadoras. Del mismo modo, no se ponderan los coeficientes para el cálculo de la fiabilidad media, pero sí para su intervalo de confianza y para el análisis de variables moderadoras. Coincide con el modelo de efecto fijo en que los resultados sólo pueden generalizarse a una población de estudios idénticos a los incluidos en el meta-análisis.

El *modelo de efectos aleatorios* parte del supuesto de que los estudios incluidos en el meta-análisis constituyen una muestra representativa (y, en sentido estricto, aleatoria) de una población mayor de estudios existentes o que podrían realizarse en el futuro, de forma que su capacidad de generalización es mayor, ya que este modelo permite generalizar los resultados a dicha superpoblación de estudios potenciales y no exactamente idénticos a los incluidos en el meta-análisis. El modelo estadístico puede formularse como $T_i = \mu_\alpha + u_i + e_i$, donde μ_α es el coeficiente alfa paramétrico medio de la superpoblación de estudios y e_i representa el error debido al muestreo de los estudios a partir de una superpoblación de estudios. Los coeficientes alfa paramétricos se asume que se distribuyen según $\alpha_i \sim N(\mu_\alpha, \tau^2)$, siendo τ^2 la varianza inter-estudios.

Dentro de este modelo se han propuesto dos modalidades. Una de ellas parte del modelo propuesto por Hunter y Schmidt (2004), según el cual los coeficientes alfa no se transforman y el factor de ponderación de dichos coeficientes en los análisis estadísticos es el tamaño muestral. La otra propuesta parte del modelo formulado por Hedges y Vevea (1998), según el cual los coeficientes alfa tendrían que ser previamente transformados para garantizar la normalidad de su distribución y estabilizar las varianzas (por ejemplo, mediante la transformación de Bonett, 2002) y, además, el factor de ponderación queda definido por la inversa de la varianza de los coeficientes transformados, siendo ésta la suma de las varianzas intra- e inter-estudios. Una modificación adicional del modelo de Hedges y Vevea (1998) ha sido propuesta por Knapp y Hartung (2003) que logra mejorar el comportamiento de las técnicas estadísticas inferenciales.

MÉTODO

Para alcanzar los objetivos se utilizó el método de caso. En concreto, se seleccionó la *Escala Yale-Brown de Obsesiones y Compulsiones* (Y-BOCS), un test elaborado por Goodman, Price, Rasmussen *et al*. (1989) para evaluar los síntomas de obsesiones y compulsiones en personas con trastorno obsesivo-compulsivo, en

personas con otros trastornos similares y también en población normal con propósitos de cribado. El test consta de 10 ítems tipo Likert (0, síntoma ausente; 4, síntoma presente) y se ha adaptado a diversas lenguas y culturas. Cada estudio empírico que reportó al menos un coeficiente alfa con los datos propios se incluyó en el meta-análisis. Para los propósitos de esta investigación, la base meta-analítica de estudios estuvo compuesta por los 46 coeficientes alfa de otros tantos estudios que aplicaron la escala Y-BOCS original. Además, para examinar el influjo de variables moderadoras se seleccionaron la desviación típica de la puntuación total del test y la población de referencia de la muestra (clínica vs. no clínica).

Sobre esta base meta-analítica se aplicaron los seis procedimientos de análisis antes descritos: (a) el método convencional basado en mínimos cuadrados ordinarios y el basado en la ponderación por la inversa de la varianza, ambos pertenecientes al *modelo de efecto fijo*, (b) el método propuesto por Bonett (2010) basado en el *modelo de coeficientes variables* y (c) tres versiones basadas en el *modelo de efectos aleatorios* (la original de Hedges y Vevea, 1998, la basada en Knapp y Hartung, 2003, y la propuesta por Hunter y Schmidt, 2004). Con cada uno de los seis métodos objeto de comparación, se calculó el coeficiente alfa medio con su intervalo de confianza y, mediante meta-regresión, se comprobó el influjo de las variables moderadoras 'desviación típica de las puntuaciones del test' y 'población de procedencia', ambas por separado. También se aplicó una meta-regresión múltiple para examinar el influjo de la variable 'población de referencia' una vez parcializado el influjo de la desviación típica de las puntuaciones.

RESULTADOS

La Tabla 1 presenta los resultados relativos al cálculo del coeficiente alfa medio y a la amplitud confidencial de su intervalo de confianza. En lo que respecta al coeficiente alfa medio, tomando como método de referencia el convencional basado en mínimos cuadrados ordinarios, no se observan grandes discrepancias entre los diferentes métodos de cálculo, siendo tan sólo de un 3,7% la mayor. En cuanto a la amplitud confidencial sí se observan importantes diferencias entre los diferentes métodos. Cabe destacar la discrepancia entre el método de Hedges y Olkin de efecto fijo, con un intervalo confidencial en torno a un 75% más estrecho que el basado en el método convencional.

Tabla 1. Comparación del coeficiente alfa medio y de la amplitud confidencial

Modelo	Método de estimación	Media	Discrep.	A.C.	Discrep.
Efectos Fijos	Convencional (OLS)	0,846	–	0,0396	–
	Hedges y Olkin (1985)	0,877	3,7%	0,0093	-76,5%
Coef. variables	Bonett (2010)	0,846	0,0%	0,0290	-26,8%
Efectos Aleatorios	Hedges y Vevea (1998)	0,863	2,0%	0,0294	-25,7%
	Knapp y Hartung (2003)	0,863	2,0%	0,0294	-25,7%
	Hunter y Schmidt (2004)	0,869	2,8%	0,0298	-24,7%

OLS: mínimos cuadrados ordinarios. A.C.: amplitud confidencial.

La Tabla 2 presenta los niveles de significación estadística y las proporciones de varianza explicada obtenidos con los diferentes métodos de meta-regresión para las variables moderadoras 'desviación típica' y 'población de referencia'. Como puede observarse, con algunas variables moderadoras los resultados de los diferentes métodos tienden a confluir (e.g., la desviación típica), mientras con otras se producen discrepancias importantes, tanto en los niveles *p* como en las estimaciones de la proporción de la varianza explicada por el moderador (e.g., la población de referencia, sola y ajustada por la desviación típica).

Tabla 2. Niveles de significación estadística y las proporciones de varianza explicada

Modelo	Método de estimación	Desviación típica			Población			Población / DT		
		Nivel *p*		R^2	Nivel *p*		R^2	Nivel *p*		R^2
Efecto fijo	Convencional	0,2779	ns	0,034	0,0412	*	0,093	0,050	*	0,105
	Hedges y Olkin	0,0111	*	0,076	2,9x10-11	***	0,171	8,7x10-13	***	0,324
Coef. var.	Bonett (2010)	0,2813	ns	–	–	*	–	0,00009	***	–
Efectos aleatorios	Hedges-Vevea	0,5452	ns	0,035	0,0098	**	0,278	0,001	***	0,278
	Knapp-Hartung	0,5921	ns	0,035	0,0258	*	0,278	0,007	**	0,278
	Hunter-Schmidt	0,4746	ns	0,055	0,1073	ns	0,169	0,114	ns	0,285

* $p < .05$. ** $p < .01$. *** $p < .001$. DT: desviación típica.

DISCUSIÓN

El propósito de este estudio fue comprobar, mediante el método de caso, cómo diferentes modelos y métodos estadísticos difieren en los resultados obtenidos cuando se aplican a un conjunto de coeficientes alfa. Nuestros resultados confirman la existencia de importantes diferencias entre los métodos. Si bien la elección del modelo estadístico por parte del meta-analista debería ser un paso previo fundamental, el hecho de que los diferentes modelos estadísticos arrojen importantes discrepancias entre ellos refuerza la necesidad de este paso.

La elección del modelo estadístico debería estar guiada por el grado de generalización que el meta-analista pretende alcanzar con sus resultados. El modelo de efectos aleatorios permite una mayor generalización, pero requiere del cumplimiento de más supuestos que los modelos de efecto fijo y de coeficientes variables. En concreto, el modelo de efectos aleatorios debería aplicarse cuando se disponga de un número razonablemente elevado de coeficientes de fiabilidad (en torno a 30 como mínimo) y se pueda garantizar que los estudios meta-analizados constituyen una muestra representativa de la superpoblación de estudios a la que se pretende generalizar los resultados.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por la Fundación Séneca de la C.A. de la Región de Murcia (Proyecto nº 08650/PHCS/08).

REFERENCIAS

Bonett, D.G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*, 335-340.

Bonett, D.G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15*, 368-385.

Borenstein, M.J., Hedges, L.V., Higgins, J.P.T. y Rothstein, H. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97-111.

Botella, J., Suero, M. y Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*, 386-397.

Goodman, W.K., Price, L.H., Rasmussen, S.A., Mazure, C., Fleischmann, R.L., Hill, C.L., Heninger, G.R. y Charney, D.S. (1989). The Yale-Brown Obsessive Compulsive Scale. I: Development, use, and reliability. *Archives of General Psychiatry, 46*, 1006-1011.

Hedges, L.V. y Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L.V. y Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Henson, R.K. y Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting «reliability generalization» studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

Hunter, J.E. y Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting errors and bias in research findings* (2ª ed.). Newbury Park, CA: Sage.

Knapp, G. y Hartung, J. (2003). Improved tests for a random-effects meta-regression with a single covariate. *Statistics in Medicine, 22*, 2693-2710.

Sánchez-Meca, J., López-Pina, J.A. y López-López, J.A. (2008). Una revisión de los estudios meta-analíticos de generalización de la fiabilidad. *Escritos de Psicología*, *1-2*, 107-118.

Sánchez-Meca, J., López-Pina, J.A. y López-López, J.A. (2009). Generalización de la fiabilidad: Un enfoque meta-analítico aplicado a la fiabilidad. *Fisioterapia*, *31*, 262-270.

Thompson, B. (Ed.) (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

META-ANALISIS DE RESULTADOS EXPERIMENTALES EXPRESADOS COMO CURVAS ROC

Juan Botella[1], Huiling Huang[1], Manuel Suero[1], Hilda Gambara[1] y Jesús Privado[2]

[1] Universidad Autónoma de Madrid

[2] Universidad Complutense de Madrid

Resumen

Hay muchos contextos de investigación experimental en los que los participantes deben emitir respuestas binarias y los datos se analizan y se expresan como curvas ROC. Se diferencian de otros en que se cuenta con un auténtico estándar de oro que permite una valoración inequívoca de cada respuesta y en que el umbral asociado a la regla de clasificación es implícito y, por tanto, probablemente sufre oscilaciones tanto inter como intra-sujetos. Aunque se han realizado síntesis meta-analíticas de este tipo de resultados, especialmente de memoria de reconocimiento, los procedimientos empleados han sido incompletos, o incluso incorrectos. No se han tenido en cuenta las dificultades derivadas de las variaciones en los tamaños de los materiales empleados o de las muestras de participantes, ni tampoco que las tasas de aciertos y falsas alarmas covarían cuando el umbral es variable. Proponemos procedimientos que superan esas dificultades.

RESULTADOS EXPERIMENTALES EXPRESADOS COMO CURVAS ROC

La Teoría de la Detección de Señales (TDS) se ha empleado mucho en psicología como modelo para acomodar marcos conceptuales explicativos de diversos procesos (Logan, 2004; Swets, Dawes y Monahan, 2000; Wickens, 2001). Entre ellos destacan una variedad de problemas en psicofísica, percepción, decisión o memoria de reconocimiento. Estos ámbitos de estudio comparten dos características que los distinguen de otros: (a) que el llamado *gold standard* es incuestionable, dado que responde a meras manipulaciones experimentales; y (b) que como el criterio de clasificación es implícito cabe esperar que sea variable.

Un buen ejemplo son los experimentos de memoria de reconocimiento. Tienen típicamente dos fases: la fase de estudio y la fase de test. En la primera los participantes son expuestos a unos materiales (a menudo palabras) con los que tienen que realizar diferentes tareas (memorizarlos explícitamente, leerlos, transformarlos, etc). En la fase de test se presentan unos materiales, de los que algunos formaban

parte de los que se presentaron en la fase de estudio (ítems *Viejos*) y otros no se habían presentado (ítems *Nuevos*). La tarea consiste en señalar los ítems que habían sido presentados en la fase de test (ítems Viejos). Llamando en términos generales señales (S) a los ítems que hay que identificar y ruidos (R) a los demás, las respuestas SI/NO generan la tabla de contingencia de la figura 1.

		Item: S	Item: R
Respuesta	Si	Acierto (A)	Falsa alarma (FA)
	No	Omisión (O)	Rechazo correcto (RC)
		N_S	N_R

Figura 1. Tabla de contingencias entre eventos y respuestas.

La información imprescindible se puede resumir en dos valores, la proporción de Aciertos ($PA = N_A/N_S$) y la proporción de Falsas Alarmas ($PFA = N_{FA}/N_R$). PA y PFA son estimaciones de las áreas que el umbral (en una supuesta variable latente de 'familiaridad') deja a su derecha en cada una de las dos distribuciones de la figura 2. Con esas dos proporciones se hacen estimaciones de los estadísticos que reflejan la capacidad para discriminar los ítems S y R y el umbral o criterio de respuesta empleado. Los indicadores más frecuentes son el par [d'; λ] o el par [A'; B''_D] (MacMillan y Creelman, 2005; Wickens, 2001). Aunque aquí nos referiremos a los primeros, nuestras argumentaciones se pueden extender a otros pares de índices que reflejan características similares de los datos.

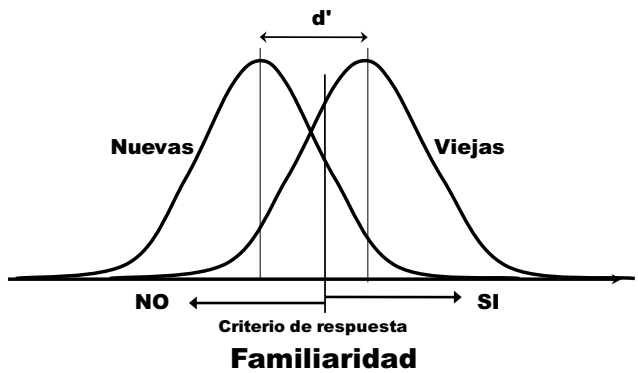


Figura 2. Distribuciones de la variable familiaridad de los ítems Señal y Ruido, con el criterio de respuesta positiva, asumiendo un modelo normal bivariado homocedástico

Como en otros campos de la psicología, a veces se hacen integraciones meta-analítica de los resultados de diversos estudios primarios. Los estudios primarios han de ser suficientemente homogéneos en cuanto a los constructos que se estudian

y los procedimientos que se emplean. Se han hecho algunos meta-análisis sobre experimentos de memoria de reconocimiento con el paradigma experimental que hemos descrito.

PROBLEMAS DE LOS META-ANÁLISIS CON DATOS DE RECONOCIMIENTO

Habitualmente, una síntesis meta-analítica incluye tanto la estimación combinada de parámetros de los que los estudios aportan estimaciones independientes como el análisis de variables moderadoras que puedan explicar parte de la variación observada en esas estimaciones (Botella y Gambara, 2002; Cooper, Hedges y Valentine, 2009; Sánchez-Meca y Botella, 2010).

En el campo de la memoria de reconocimiento se han publicado algunos estudios que se autocalifican con la etiqueta «meta-análisis», en los que se hace un análisis de los resultados de experimentos con el diseño que hemos expuesto (Donaldson, 1996; Gardiner, Ramponi y Richardson-Klavehn, 2002). En estos estudios se elabora una tabla con las proporciones básicas de cada estudio, PA y PFA, así como las estimaciones de los parámetros del modelo asumido. Por ejemplo, asumiendo un modelo normal bivariado homocedástico, la tabla podría incluir las estimaciones de *d'* y λ de cada estudio primario (tabla 1).

Tabla 1. Tabla de resultados resumen de K estudios

Estudio	PA	PFA	d'	λ
1	PA_1	PFA_1	d'_1	λ_1
2	PA_2	PFA_2	d'_2	λ_2
-	-	-	-	-
j	PA_j	PFA_j	d'_j	λ_j
-	-	-	-	-
k	PA_K	PFA_K	d'_k	λ_k
Media	$\overline{PA}$	$\overline{PFA}$	$\overline{d'}$	$\overline{\lambda}$

Se han empleado dos procedimientos para tratar los datos en forma de promedio simple. En el primero, *Promedio1*, se hallan las medias aritméticas de las proporciones PA y PFA y con estos se estiman *d'* y λ. En el segundo, *Promedio2*, se

obtienen las estimaciones de los parámetros (por ejemplo, *d'* y λ) para cada estudio y posteriormente se halla su media aritmética. Los dos estudios mencionados arriba emplean sobre todo *Promedio2*, aunque también argumentan mediante *Promedio1*. También se obtienen correlaciones entre las proporciones o entre los estadísticos observados.

Estos dos procedimientos son inadecuados porque no tienen en cuenta algunas circunstancias que son habitualmente superadas con los procedimientos más usuales del meta-análisis. Vamos a discutir las tres limitaciones principales: variaciones entre estudios en el número de ítems, covariación entre PA y PFA, y variaciones entre estudios en el número de participantes.

VARIACIONES EN EL NÚMERO DE ÍTEMS

Cada uno de los *K* estudios emplea un número diferente de ítems, N_{S_j} y N_{R_j}. Representando por π_A la probabilidad de un acierto (probabilidad de responder SI dado un ítem S), la proporción empírica de aciertos del estudio *j*, $PA_j = A_j/N_{S_j}$, es un estimador de π_A (la proporción empírica de falsas alarmas, $PFA_j = FA_j/N_{R_j}$, es un estimador de π_{FA}). Los estimadores que proporciona cada estudio son más precisos cuanto mayor es el número de ítems (N_{S_j} y N_{R_j}). Cuando se dispone de varios estimadores de un mismo parámetro, como *K* valores de PA_j (o los *K* valores de PFA_j) el estimador combinado de mínima varianza es el promedio ponderado de esas estimaciones, siendo los pesos iguales a los inversos de sus varianzas (Hedges y Olkin, 1985). Si en el estudio *j* se emplean N_{S_j} ítems, la varianza de PA_j y su peso para la estimación combinada (w_j) son,

$$Var(PA_j) = \frac{\pi_A \cdot (1-\pi_A)}{N_{S_j}} \qquad [1]$$

$$w_j = 1 \Big/ Var(PA_j) = \frac{N_{S_j}}{\pi_A \cdot (1-\pi_A)} \qquad [2]$$

Por tanto, el estimador combinado de mínima varianza de una proporción (de aciertos o de falsas alarmas) es,

$$\hat{\pi}_\bullet = \frac{\sum_{j=1}^{K} w_j \cdot P\bullet_j}{\sum_{j=1}^{K} w_j} \qquad [3]$$

donde $\bullet$ es A ó FA.

Según esto, los procedimientos *Promedio1* y *Promedio2* son inadecuados porque no tienen en cuenta este aspecto de los datos. Las estimaciones obtenidas con [3] son más adecuadas que la media simple. Naturalmente, en el procedimiento *promedio2* los estimadores de los parámetros se deben ponderar con las varianzas de esos estimadores, como veremos más adelante.

COVARIACIÓN ENTRE PA Y PFA

En el procedimiento *Promedio1* se hacen estimaciones independientes de PA y PFA. Pero cuando el umbral de clasificación es implícito, como ocurre en este tipo de experimentos, son esperables ciertas oscilaciones en el valor del criterio, lo que implica una covariación positiva entre π_A y π_{FA}. No tenerlo en cuenta puede llevar a errores graves, tales como una estimación errónea de *d'* (véase Macmillan y Creelman, 2005).

El procedimiento *Promedio2* sí tiene en cuenta indirectamente el balance entre PA y PFA, resolviendo el problema de la covariación. Sin embargo, los meta-análisis realizados con este procedimiento no tienen en cuenta la cuestión de la ponderación.

Los análisis estadísticos sobre las estimaciones de los parámetros se deberían realizar de nuevo mediante ponderaciones basadas en los inversos de sus varianzas. Supongamos que asumimos el par [*d';λ*] (Wickens, 2001), que reflejan la capacidad para discriminar los dos estados (*d'*) y el criterio de clasificación (*λ*). Una aproximación a la varianza del índice del criterio (*λ*) es (Gourevitch y Galanter, 1967):

$$\mathrm{Var}\left(\hat{\lambda}\right) = \frac{\mathrm{Var(PFA)}}{[\varphi(\hat{\lambda})]^2} \qquad [4]$$

donde Var(*PFA*) se define en [1] y $\varphi\left(\hat{\lambda}\right)$ es el valor de la función de densidad de la distribución normal de R en *λ*. La varianza aproximada de *d'* es (Gourevitch y Galanter, 1967),

$$\mathrm{Var}(\hat{\mathrm{d}}') = \frac{\mathrm{Var(PFA)}}{[\varphi(\hat{\lambda})]^2} + \frac{\mathrm{Var(PA)}}{[\varphi(\hat{\mathrm{d}}'-\hat{\lambda})]^2} \qquad [5]$$

donde Var(PA) y $\varphi(\hat{\mathrm{d}}'-\hat{\lambda})$ son análogos a los elementos ya definidos (esta fórmula asume que *d'* y *λ* son independientes, algo que no siempre se cumple). Se pueden aplicar los procedimientos habituales en meta-análisis con los inversos de estas varianzas como pesos,

$$\hat{\mathrm{d}}' = \frac{\sum_{j=1}^{K} \mathrm{w_j} \cdot \mathrm{d'_j}}{\sum_{j=1}^{K} \mathrm{w_j}} \qquad [6]$$

$$\left(\mathrm{w_j} = \frac{1}{\mathrm{Var(d'_j)}}\right)$$

$$\hat{\lambda} = \frac{\sum_{j=1}^{K} \mathrm{w_j} \cdot \lambda_j}{\sum_{j=1}^{K} \mathrm{w_j}} \qquad [7]$$

$$\left(\mathrm{w_j} = \frac{1}{\mathrm{Var}(\lambda_j)}\right)$$

Este procedimiento tiene en cuenta los dos problemas que venimos mencionando y los resuelve adecuadamente. Hasta donde sabemos nosotros, no ha sido utilizado con resultados de experimentos como los descritos anteriormente.

VARIACIONES EN EL NÚMERO DE PARTICIPANTES

Un problema adicional es que la varianza de los estimadores no sólo es función del número de ensayos, sino también del número de participantes. En las tablas que resumen los estudios se incluyen proporciones de aciertos y falsas alarmas, pero estas proporciones son los promedios de los participantes de cada experimento. Los estudios no proporcionan los datos individuales, sino los promedios de los M_j participantes del estudio *j* en cada condición experimental.

Para un observador que realiza N_{S_j} ensayos con ítems S la varianza de la proporción de aciertos es la que se expresa en [1]. Los estudios primarios proporcionan promedios de esas proporciones,

$$\overline{PA}_j = \frac{\sum_{i=1}^{M} PA_{ij}}{N_{S_i}} \qquad [8]$$

Asumiendo el supuesto, bastante razonable, de que las observaciones de los participantes son independientes, la varianza de este estadístico es la suma de las varianzas de los propios participantes multiplicada por $(1/M_j)^2$. Supongamos que la probabilidad de emitir una respuesta positiva (π_A y π_{FA}) es constante en todos los participantes en una misma condición experimental; entonces para los aciertos,

$$Var(\overline{PA}_j) = \frac{1}{M_j^2} \cdot \sum_{i=1}^{M} Var(PA_j) = \frac{1}{M_j^2} \cdot M_j \cdot \frac{\pi_A \cdot (1-\pi_A)}{N_{S_j}} = \frac{\pi_A \cdot (1-\pi_A)}{M_j \cdot N_{S_j}} \qquad [9]$$

Es decir, la varianza sería M_j veces menor que la de cada observador individual. Por tanto, los pesos adecuados cuando los estudios aportan promedios de las proporciones de aciertos son (para las falsas alarmas las fórmulas son similares),

$$w_j = 1 \Big/ Var(PA_j) = \frac{M_j \cdot N_{S_j}}{\pi_A \cdot (1-\pi_A)} \qquad [10]$$

Las fórmulas [4] y [5] quedarían,

$$\mathrm{Var}\left(\hat{\lambda}\right) = \frac{\pi_{FA} \cdot (1-\pi_{FA})}{M_j \cdot N_{R_j} \cdot [\varphi(\hat{\lambda})]^2} \qquad [11]$$

$$\mathrm{Var}(\hat{d}') = \mathrm{Var}(\hat{\lambda}) + \frac{\pi_A \cdot (1-\pi_A)}{M_j \cdot N_{S_j} \cdot [\varphi(\hat{d}'-\hat{\lambda})]^2} \qquad [12]$$

Sin embargo, hemos asumido que las probabilidades de emitir un acierto y una falsa alarma, π_A y π_{FA}, son contantes para todos los individuos y estudios. Este es un supuesto poco creíble. Es más verosímil que esas probabilidades muestren diferencias entre los individuos, pero también entre los estudios, aunque se hagan manipulaciones experimentales similares. En meta-análisis es habitual incorporar esta característica mediante el ajuste de modelos de efectos aleatorios, en lugar de modelos de efectos fijos.

En un modelo de efectos fijos se asume que cada estudio implica una muestra procedente de una variable aleatoria con distribución única. En un modelo de efectos aleatorios se asume que los valores de los parámetros de cada estudio proceden de una distribución que se puede describir con un hiperparámetro. Por ejemplo, se puede asumir que la probabilidad de acierto en el estudio *j* es π_{A_j}. Los potenciales estudios con características homogéneas tendrán unas probabilidades paramétricas que seguirían una determinada distribución con valor esperado π_{A_j} y varianza $\sigma^2_{\pi_A}$. Al final, las proporciones de aciertos observadas tendrán dos fuentes de variación. Por un lado, la propia de cualquier proceso de muestreo de observaciones empíricas, que se describe en la fórmula [1]; por otro, la debida a las variaciones en π_A debidas tanto al muestreo de observadores como a las variaciones entre estudios. Representando por $v_{d'}$ y v_λ a las variaciones extra en *d'* y λ debidas a estas fuentes, el modelo de efectos aleatorios es,

$$\hat{d}' = \frac{\sum_{j=1}^{K} w_j \cdot \hat{d}'_j}{\sum_{j=1}^{K} w_j} \qquad [13]$$

$$\left(w_j = \frac{1}{\mathrm{Var}(\hat{d}'_j) + v_{d'}} \right)$$

$$\hat{\lambda} = \frac{\sum_{j=1}^{K} w_j \cdot \hat{\lambda}_j}{\sum_{j=1}^{K} w_j} \qquad [14]$$

$$\left(w_j = \frac{1}{\mathrm{Var}(\hat{\lambda}_j) + v_\lambda} \right)$$

El ajuste de este modelo exigirá la estimación de las varianzas específicas de los parámetros de discriminación y criterio de respuesta, $v_{d'}$ y v_λ, que caracterizan a los modelos de efectos aleatorios (Hedges y Vevea, 1998). A diferencia de lo que ocurre en otros contextos, aquí estos parámetros recogen simultáneamente la variación global entre los π_A que se produce tanto como variaciones entre los participantes de un experimento como las variaciones entre los experimentos debidos a las diferentes características de procedimiento, materiales, etc.

CONCLUSIONES

La forma como se ha venido analizando los datos de memoria de reconocimiento en algunos autodenominados 'meta-análisis' es incorrecta. Aunque las diferencias entre esos procedimientos y otros más adecuados puedan ser cuantitativamente pequeñas en algunos casos, es probable que con otras bases de datos no sea así. Los procedimientos empleados se deben proteger frente a los efectos de diversos factores que afectan a las estimaciones.

Un procedimiento correcto para tratar este tipo de datos debe tener en cuenta: (a) el número de ítems en el que se basan las estimaciones de PA y PFA de cada estudio; (b) La covariación entre PA y PFA, típica de situaciones en las que el criterio es implícito; y (c) el tamaño de las muestras. Los estudios con estimaciones medias basadas en más participantes deberían tener más peso en las estimaciones combinadas.

El procedimiento que hemos propuesto cumple con estas condiciones. Es el modelo de distribuciones normales homocedásticas, con ponderación de d' y λ por el inverso de su varianza (teniendo en cuenta el número de ítems y de participantes). Además, es coherente con el marco conceptual de numerosos campos de la psicología. Aunque exige asumir el supuesto de normalidad, éste no es más arriesgado que el de la distribución logística, asumido en otros procedimientos. Hacen falta estudios de simulación que nos permitan valorar el impacto que tienen los procedimientos expuestos en la estimación de los parámetros.

NOTA DE LOS AUTORES

Investigación financiada con el proyecto PSI2009-12071, Ministerio de Ciencia y Tecnología

REFERENCIAS

Botella, J. y Gambara, H. (2002). Qué es el meta-análisis. Madrid: Biblioteca Nueva.

Cooper, H., Hedges, L. V. y Valentine, J. C. (2009). *The handbook of research synthesis and Meta-analysis, 2nd ed*. Nueva York: Russell Sage Foundation.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, 24 (4),523-533.

Gardiner, J. M., Ramponi, C., Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory*, 10 (2), 83–98.

Gourevitch, V. y Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, 32, 25-33.

Hedges L. V., & Olkin, I. (1985). *Statistical methods of meta-analysis*. Orlando, Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Logan, G. D. (2004). Cumulative Progress in Formal Theories of Attention. *Annual Review of Psychology*, 55, 207-234.

MacMillan, N. A. y Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2ª ed.). Mahwah, NJ: Erlbaum.

Sánchez-Meca, J. y Botella, J. (2002). Revisiones sistemáticas y meta-análisis: herramientas para la práctica profesional. *Papeles del Psicólogo*, 31(1), 7-17.

Swets, J. A., Dawes, R. M. y Monahan, J. (2000). Psychological Science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1(1), 1-26.

Wickens, T. D. (2001). *Elementary signal detection theory*. Nueva York: Oxford University Press.

META-ANÁLISIS DE COEFICIENTES ALFA: ¿ES ÚTIL LA FÓRMULA KR-21?

Julio Sánchez-Meca, José A. López-Pina y José A. López-López
Universidad de Murcia
Correo electrónico: jsmeca@um.es

Resumen

Uno de los principales problemas de los estudios de generalización de la fiabilidad es el fenómeno de la inducción de la fiabilidad, según el cual, la mayoría de los estudios empíricos que aplican un determinado instrumento de medida no suelen reportar una estimación de la fiabilidad con los propios datos de la muestra, sino que inducen la fiabilidad a partir de algún estudio previo (generalmente, mediante la obtenida en el estudio original de validación del test). Se estima que menos del 20% de los estudios empíricos reportan la fiabilidad de los tests empleados en la muestra. Cuando el test está compuesto por ítems dicotómicos, es posible incrementar la base de estudios meta-analizables aunque éstos no reporten el coeficiente alfa, mediante la fórmula *KR*-21, siempre que el estudio reporte la media y la desviación típica de la puntuación total del test. Mediante su aplicación a un ejemplo concreto, en este trabajo se examina la utilidad de la fórmula *KR*-21 para incrementar la base de estudios meta-analizables. Específicamente, se analiza el sesgo de *KR*-21 como estimador del coeficiente alfa, la varianza compartida por ambos y la utilidad de *KR*-21 para optimizar el análisis de variables moderadoras de los coeficientes de fiabilidad.

Una práctica habitual en las investigaciones psicológicas que aplican tests a muestras de sujetos es inducir la fiabilidad del test a partir de la obtenida en alguna aplicación previa del mismo, generalmente en el estudio de validación del test. Esta práctica, conocida como 'inducción de la fiabilidad', asume que la fiabilidad es una propiedad inmutable del test y que, por tanto, no varía en diferentes aplicaciones del mismo (Vacha-Haase, Kogan y Thompson, 2000). Sin embargo, tal y como se calculan los coeficientes de fiabilidad desde la teoría clásica de los tests, la fiabilidad es una propiedad que varía de una administración a otra del instrumento, ya que depende de la composición y variabilidad de la muestra de participantes sobre la que el test se aplica. Una de esas características es la variabilidad de las puntuaciones del test en la muestra objeto de estudio (Botella, Suero y Gambara, 2010). Es por ello que frases tan comunes como 'la fiabilidad del test es 0,80' inducen a error, ya que la fiabilidad debe considerarse una propiedad de las puntuaciones obtenidas en una aplicación concreta del test, y no como una propiedad inmutable de éste (Crocker y Algina, 1986).

Dado que la fiabilidad varía con las administraciones del test, el meta-análisis se convierte en una metodología idónea para examinar la fiabilidad media de las puntuaciones del test a lo largo de sus diversas aplicaciones y para explorar qué características de las muestras de participantes y del contexto de aplicación del test son capaces de explicar la variabilidad de un conjunto de coeficientes de fiabilidad. Con este fin, Vacha-Haase (1998) acuñó el término *generalización de la fiabilidad* (GF) para referirse a este tipo de meta-análisis psicométrico (Henson y Thompson, 2002; Sánchez-Meca y López-Pina, 2008; Sánchez-Meca, López-Pina y López-López, 2008, 2009; Thompson, 2003).

Para poder realizar un estudio de GF sobre un determinado test es preciso disponer de un conjunto de estudios primarios que hayan aplicado el mismo y que reporten un coeficiente de fiabilidad obtenido con los propios datos. Sin embargo, debido a la práctica generalizada en los investigadores de inducir la fiabilidad, son escasos los estudios que reportan una estimación propia de la fiabilidad. En una revisión sistemática recientemente realizada por nuestro equipo recabando información de 84 estudios de GF realizados hasta la fecha, hemos encontrado que, en promedio, tan sólo el 34% de los estudios que han aplicado un test suelen reportar un coeficiente de fiabilidad de sus puntuaciones (Sánchez-Meca, López-Pina y López-López, 2011, mayo-junio).

Como consecuencia de esta desaconsejable práctica de inducir la fiabilidad de las puntuaciones del test, la mayoría de los estudios primarios que lo han aplicado tienen que excluirse del estudio de GF al no aportar un coeficiente de fiabilidad propio. Es por ello que se convierte en una tarea prioritaria encontrar estrategias que permitan incrementar la base de estudios meta-analizables, aun cuando no reporten un coeficiente de fiabilidad. El tipo de fiabilidad más habitualmente reportado en los estudios es la basada en la consistencia interna de sus puntuaciones, en concreto, mediante el cálculo del coeficiente alfa. Cuando el test está formado por ítems dicotómicos y con un nivel similar de dificultad, la fórmula *KR*-21 constituye una buena estimación del coeficiente alfa. Para calcular el coeficiente *KR*-21 sólo se precisan el número de ítems del test, *J*, la media y la desviación típica de las puntuaciones totales del test ($\overline{X}$ y S_x) según la fórmula:

$$KR-21=\frac{J}{J-1}\left[1-\frac{\overline{X}(J-\overline{X})}{JS_x^2}\right].$$

Así pues, aquellos estudios que no aportan una estimación propia de la fiabilidad pero reportan la media y la desviación típica de las puntuaciones del test, serían susceptibles de ser incluidos en el estudio (Henson, Kogan y Vacha-Haase, 2001). No obstante, investigaciones previas han constatado la existencia de un ligero sesgo negativo de la fórmula *KR*-21 como estimador del coeficiente alfa, que se sitúa en torno a 0,03 como máximo cuando el coeficiente alfa es superior a 0,60 (Betebenner y Hopkins, 1997). Así, en su estudio de GF sobre el *Inventario de Autoestima de Coopersmith* (CSI), Lane, White y Henson (2001) encontraron que la fórmula KR-21 presentó un sesgo negativo de tan sólo 0,022 respecto del coeficiente

alfa. Y Kieffer y Reese (2002), en su estudio de GF sobre la *Escala de Depresión Geriátrica* (DGS) obtuvieron un sesgo negativo de 0,05.

El propósito de este estudio fue, mediante su aplicación a un caso real, comprobar la utilidad de la fórmula *KR*-21 como un sustituto del coeficiente alfa cuando éste no es reportado en el estudio, con objeto de incrementar la base de estudios del meta-análisis. Además, estimamos el sesgo negativo de la fórmula *KR*-21 como estimador del coeficiente alfa y la varianza que comparten ambos.

MÉTODO

Para alcanzar los objetivos se utilizó el método de caso. En concreto, se seleccionó el Inventario de Obsesiones y Compulsiones de Maudsley (MOCI), un test elaborado por Hodgson y Rachman (1977) para evaluar los síntomas de obsesiones y compulsiones en personas con trastorno obsesivo-compulsivo, en personas con otros trastornos similares y también en población normal con propósitos de cribado. El test consta de 30 ítems dicotómicos y se ha adaptado a diversas lenguas y culturas (Sánchez-Meca, López-Pina, López-López, Marín-Martínez, Rosa-Alcázar y Gómez-Conesa, 2011). Los detalles del meta-análisis de GF de este Inventario pueden consultarse en Sánchez-Meca *et al.* (2011). Cada estudio empírico que reportó al menos un coeficiente alfa con los datos propios se incluyó en el meta-análisis. Además, aquellos estudios que aportaron el número de ítems del test, la media y la desviación típica de la puntuación total del test también se incluyeron en el meta-análisis para obtener una estimación del coeficiente alfa mediante la fórmula *KR*-21. Con este propósito, antes de aplicar la fórmula *KR*-21 se calculó la varianza no corregida de las puntuaciones del test, S_x^2, a partir de la varianza corregida o insesgada que es reportada habitualmente en los estudios, S_{n-1}^2, mediante $S_x^2 = S_{n-1}^2\left[(n-1)/n\right]$.

Con objeto de normalizar la distribución de los coeficientes de fiabilidad y estabilizar sus varianzas, éstos fueron previamente transformados mediante la fórmula propuesta por Bonett (2002): $T = Ln(1-|r|)$, siendo r el coeficiente de fiabilidad (ya sea alfa o *KR*-21) y *Ln* el logaritmo natural. Posteriormente, con objeto de facilitar la interpretación de los resultados, los valores T fueron retransformados a la métrica del coeficiente de fiabilidad mediante: $r = 1 - e^T$. Los análisis estadísticos para obtener la fiabilidad media y su intervalo de confianza se realizaron asumiendo un modelo de efectos aleatorios, según el cual cada coeficiente de fiabilidad transformado se ponderó por la inversa de su varianza, definida ésta como la suma de la varianza intra-estudio y la varianza inter-estudios (Borenstein, Hedges, Higgins y Rothstein, 2009).

RESULTADOS

De los 418 estudios empíricos que aplicaron el test MOCI, tan sólo 40 estudios reportaron un coeficiente alfa (9,5%). De los 378 estudios que indujeron la fiabili-

dad fue posible calcular la fórmula *KR*-21 en 249 estudios, lo que permitió incrementar dramáticamente la muestra de coeficientes de fiabilidad de 40 a 289 (69,1%).

El segundo objetivo de nuestro estudio fue comprobar el sesgo de *KR*-21 como estimador del coeficiente alfa. Para ello, seleccionamos aquellos estudios que reportaron un coeficiente alfa y que nos permitieron calcular la fórmula *KR*-21 al aportar también la media y la desviación típica de las puntuaciones del test. De los 289 estudios meta-analizados, 31 de ellos aportaban esta información, lo cual nos permitió obtener la fiabilidad media con los coeficientes alfa y con la fórmula *KR*-21. La Tabla 1 presenta los resultados. El sesgo negativo obtenido por *KR*-21 respecto del coeficiente alfa en nuestro estudio fue de 0,024, un valor muy próximo al obtenido por Lane *et al.* (2001), algo inferior al obtenido por Kieffer y Reese (2002) y ligeramente por debajo de la predicción hecha por Betebenner y Hopkins (1997). No obstante, se observaron diferencias estadísticamente significativas entre los dos coeficientes medios [$T(30) = 2{,}746$, $p = .010$; $d = 0{,}49$].

Tabla 1. Análisis del sesgo de la fórmula KR-21 como estimador del coeficiente alfa

Coeficiente	*k*	Mín.	Máx.	Media	I. C. al 95%		*Q*	*p*	I^2
					Li	Ls			
Alfa	31	0,61	0,86	0,761	0,743	0,778	150,08	< .001	80,0%
KR-21	31	0,61	0,85	0,737	0,718	0,756	138,86	< .001	78,4%
			Diferencia:	**-0,024**					
			Lane *et al*. (2001):	**-0,022**					
			Kieffer y Reese (2002):	**-0,050**					
			Betebener y Hopkins (1997)	≤ **|0,030|**					

Nuestro último propósito era examinar la varianza compartida por el coeficiente alfa y las estimaciones mediante *KR*-21 con objeto de fundamentar nuestra propuesta de utilizar *KR*-21 en los estudios de GF. Para ello, aplicamos un modelo de regresión lineal simple sobre las 31 parejas de coeficientes alfa y *KR*-21, tomando como predictor *KR*-21 y los coeficientes alfa como variable dependiente. La Tabla 2 presenta los resultados, así como los obtenidos por Lane *et al*. (2001) y Betebenner y Hopkins (1997). Puede observarse cómo las pendientes de las ecuaciones de regresión en los tres estudios son similares, situándose en torno a 0,75, así como las intercepciones. Además, nuestro modelo de regresión presentó un 47,8% de varianza compartida por los dos coeficientes, lo que indica la existencia de una fuerte covariación entre ambos coeficientes.

Tabla 2. Regresión del coeficiente alfa a partir de KR-21

Estudio	*k*	Ecuación de regresión	r_{xy}	R^2
MOCI	31	α' = 0,222 + 0,730**KR*-21	0,691	0,478
Lane et al. (2001)	34	α' = 0,188 + 0,778**KR*-21	0,902	0,814
Betebenner y Hopkins (1997)	668	α' = 0,115 + 0,883**KR*-21	–	–

DISCUSIÓN

El objeto de nuestro estudio fue demostrar la utilidad de la fórmula *KR*-21 como estimador del coeficiente alfa para incrementar el número de estudios meta-analizables en un estudio de GF sobre un test compuesto por ítems dicotómicos de similar dificultad. Con este propósito, pretendíamos replicar los resultados de los estudios de Lane *et al*. (2001) y Kieffer y Reese (2002) con un estudio de GF del test MOCI (Sánchez-Meca et al., 2011). Nuestros resultados ponen de manifiesto la gran utilidad que ofrece la fórmula *KR*-21 para incrementar la base de estudios de un meta-análisis de GF. Además, el sesgo negativo de *KR*-21 como estimador del coeficiente alfa fue inferior al predicho por Betebenner y Hopkins, si bien se obtuvieron diferencias estadísticamente significativas entre ambos promedios. La varianza compartida por ambos coeficientes fue de magnitud alta.

Nuestros resultados avalan la utilidad de la fórmula *KR*-21 para incrementar la base de estudios meta-analizables en un estudio de GF. No obstante, siempre que se pretenda ampliar la base de estudios mediante esta estrategia, conviene comprobar previamente el sesgo negativo exhibido por *KR*-21 en comparación con las estimaciones directas del coeficiente alfa, así como su varianza compartida. Si el sesgo se sitúa por debajo de 0,03 y la varianza compartida es razonablemente alta, estaría perfectamente justificado utilizar *KR*-21 para estimar la fiabilidad media de las puntuaciones del test a partir de una base de estudios mayor que si sólo utilizamos los coeficientes alfa directamente reportados en los estudios empíricos. Así mismo, esta estrategia nos permitirá examinar con una base de estudios mayor las características de los estudios, o variables moderadoras, que mejor pueden dar cuenta de la variabilidad exhibida por los coeficientes de fiabilidad.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por la Fundación Séneca de la C.A. de la Región de Murcia (Proyecto nº 08650/PHCS/08).

REFERENCIAS

Betebenner, D. y Hopkins, K.D. (1997). *Estimating Cronbach's alpha from KR-21*. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (Citado en Lane *et al.,* 2002, pp. 687-688.)

Bonett, D.G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*, 335-340.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. y Rothstein, H.R. (2009). *Introduction to metaanalysis*. Chichester, UK: Wiley.

Botella, J., Suero, M. y Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*, 386-397.

Crocker, L. y Algina, J. (1986). *Introduction to classical and modern test theory*. Nueva York: Holt, Rinehart, & Winston.

Henson, R.K., Kogan, L.R. y Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*, 404-420.

Henson, R.K. y Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting «reliability generalization» studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

Hodgson, R.J. y Rachman, S. (1977). Obsessional-compulsive complaints. *Behaviour Research and Therapy, 15*, 389-395.

Kieffer, K.M. y Reese, R.J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement, 62*, 969-994.

Lane, G.G., White, A.E. y Henson, R.K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study on the Coopersmith Self-esteem Inventory. *Educational and Psychological Measurement, 62*, 685-711.

Sánchez-Meca, J. y López-Pina, J.A. (2008). El enfoque meta-analítico de generalización de la fiabilidad. *Acción Psicológica*, *5*, 37-64.

Sánchez-Meca, J., López-Pina, J.A. y López-López, J.A. (2008). Una revisión de los estudios meta-analíticos de generalización de la fiabilidad. *Escritos de Psicología*, *1-2*, 107-118.

Sánchez-Meca, J., López-Pina, J.A. y López-López, J.A. (2009). Generalización de la fiabilidad: Un enfoque meta-analítico aplicado a la fiabilidad. *Fisioterapia*, 31, 262-270.

Sánchez-Meca, J., López-Pina, J.A. y López-López, J.A. (2011, mayo-junio). *El problema de la inducción de la fiabilidad de los tests en la investigación psicológica: Una revisión sistemática*. Comunicación presentada al VIII Foro sobre la Evaluación de la Calidad de la Educación Superior y la Investigación (FECIES), Santander.

Sánchez-Meca, J., López-Pina, J.A., López-López, J.A., Marín-Martínez, F., Rosa-Alcázar, A.I. y Gómez-Conesa, A. (2011). The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology, 11*, 473-493.

Thompson, B. (Ed.) (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

Vacha-Haase, T., Kogan, L.R. y Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals. *Educational and Psychological Measurement, 60*, 509-522.

EVALUACIÓN DE LA CALIDAD EN ORGANIZACIONES DE SERVICIOS

Coordinadora: Verónica Morales
Universidad de Málaga

La evaluación de la calidad en organizaciones de servicios es una realidad imperante en las últimas décadas. El interés no ha parado de crecer, muchas de las organizaciones a través de sus directrices, están intentando orientarse sobre aquellos aspectos que deben ser evaluados a través de procesos de acreditación y normalización de la calidad, hacia el camino de excelencia. Se abre por tanto, una posibilidad real de abrir líneas de investigación en pro de la construcción de herramientas de medida fiables, válidas, precisas y pragmáticas, de fácil aplicación para la evaluación de la calidad en distintas organizaciones de servicios en los diferentes ámbitos. Utilizando el marco general de nuestra intervención, intentamos a través de éste Simposium, transmitir la idea de que el servicio es susceptible de ser conceptualizado, definido, medido y mejorada su implementación, utilizando para ello el feedback de todo el proceso descrito. La necesidad apremiante requerida en la situación actual de crisis, es desarrollar en las organizaciones de servicio un profundo sentido de la necesidad de evaluar la calidad como un requisito necesario para mejorar e incrementar su afianzamiento en la sociedad, en la fidelización y la permanencia de sus usuarios. Para conseguir este objetivo, es importante estimular y suministrar tanto a los gerentes como al personal, sistemas de evaluación, herramientas con garantías metodológicas, de cómoda aplicación que faciliten sus trabajos, para desarrollar y consolidar una cultura interna centrada en el cumplimiento de objetivos hacia una mejora continua de la calidad de servicio, constituyendo una clave genuina hacia el camino de la excelencia. Este Simposium quiere ser ejemplo de la necesaria interdisciplinariedad en éste ámbito, reuniendo a diversos investigadores de distintas universidades y de distintas disciplinas, los cuales, han podido aunar sus esfuerzos y conocimientos, para contribuir en la evolución y desarrollo de la investigación en este área. El principal objetivo en este Simposium es dar a conocer la construcción de distintos sistemas para evaluar la calidad en diversos ámbitos de intervención, ya sea en el ámbito de la actividad física y del deporte, en el ámbito del voluntariado o en el ámbito sanitario. Todas ellas aportan sus conocimientos diferenciadores e innovadores, de gran interés y sin duda, necesarias para la mejora de la calidad del servicio y la relación entre sus personas. A través de distintas comunicaciones, fruto de los trabajos realizados en los últimos años en diferentes organizaciones de servicios, tales como la gestión de servicios deportivos, clubs de golf, voluntariado deportivo, voluntariado medioambiental, en

centros de atención infantil temprana, etc., pretendemos promover, la utilización de estrategias de investigación, útiles en la práctica profesional, fáciles de aplicar e intentar un acercamiento hacia la mejora de la Calidad en la Gestión de organizaciones de servicios.

PALABRAS CLAVE: Calidad percibida, Calidad de servicio, Servicios deportivos, Gestión de la calidad, Atención temprana, Voluntariado deportivo, Voluntariado medioambiental, Gestión de la calidad en clubes de Golf.

LA PERCEPCIÓN DE LA CALIDAD EN ORGANIZACIONES DE SERVICIOS DEPORTIVOS

Verónica Morales Sánchez y Pablo Gálvez Ruíz
Universidad de Málaga

Resumen

En la presente investigación se administró el Cuestionario de Evaluación de la Calidad Percibida en Servicios Deportivos (*CECASDEP*) en sus dos versiones (*v.1.0*: 71 ítems; *v.2.0*; 51 ítems), siendo el objetivo averiguar si se mantiene su estructura factorial, utilizándose dos muestras diferentes compuestas por 110 participantes para la *v.1.0* y 537 para la *v.2.0* respectivamente. Se realizó un análisis factorial exploratorio para comprobar las propiedades psicométricas, resultando la medida de adecuación muestral satisfactoria en todas las escalas (KMO >.70) y la prueba de esfericidad de Bartlett estadísticamente significativa. Las soluciones factoriales explican el 65% (*v.1.0*) y el 55% (*v.2.0*) de la variancia, siendo el coeficiente alfa de consistencia interna adecuado en todas las escalas (α>.70). Estos resultados muestran evidencia de la estabilidad de la estructura factorial del instrumento de medida, aunque necesitan ser confirmados mediante un análisis factorial confirmatorio para evaluar su ajuste.

Las características particulares por las que atraviesa actualmente la actividad física y el deporte han supuesto, para los centros deportivos, un considerable aumento tanto de la oferta de programas de actividad física como en el número de usuarios/as. Así, en los últimos años hemos asistido a una evolución muy positiva en torno a los servicios deportivos, comprobando que la demanda ha derivado hacia una finalidad más lúdica, recreativa o de salud y donde la atención a las necesidades de los/as usuarios/as representa un aspecto esencial a la hora de establecer un modelo de gestión.

En la búsqueda tanto de la excelencia empresarial como de una alta rentabilidad, se hace necesaria la adaptación y la multidisciplinariedad de las instalaciones y espacios deportivos existentes con el objetivo de conseguir la máxima funcionalidad, la oferta de adecuados programas de actividad física en función de la demanda específica, así como también una óptima relación de la organización con los/as usuarios/as tratando de asegurar una alta tasa de fidelización (Morales Sánchez y Gálvez, 2011).

De esta forma, la prestación de servicios deportivos de calidad representa una de las estrategias más empleadas en la actualidad, por lo que el estudio de la calidad percibida por parte de los/as usuarios/as, en relación con el servicio deportivo que

reciben, supone una de las áreas de análisis que cuentan con mayor protagonismo en la gestión deportiva actual, siendo además necesario para alcanzar el éxito, según Morales et al. (2005) el establecimiento de un plan de calidad con una adecuada optimización de los recursos, la reducción de costes y una mejora continua.

Pero la preocupación por conseguir estándares de calidad en la prestación de servicios deportivos ha generado una necesidad de evaluarlos. En la última década, son numerosas las investigaciones que han utilizado adaptaciones de la herramienta *SERVQUAL* (Parasuraman et al., 1988) para la evaluación de los servicios deportivos (Barrera y Reyes, 2003; Morales Sánchez, 2003; Morales Sánchez et al., 2004, 2005, 2009; Salvador, 2005), siendo notables en la actualidad en la literatura científica las investigaciones que tienen como objetivo evaluar, mediante herramientas diseñadas específicamente, diferentes tipos de servicios deportivos o aspectos concretos del mismo, como los centros fitness (Afthinos et al., 2005; Marmol et al., 2010), los servicios náuticos (Calabuig et al., 2008), la industria del deporte recreacional o las organizaciones deportivas (Ko y Pastore, 2005; Nuviala et al., 2010; Rial et al., 2010), los practicantes de spinning (Sanz et al., 2005), los programas de actividad física (Hernández Mendo, 2001) o el servicio prestado en clubes de golf (Serrano et al., 2010), entre otros.

En este sentido, la herramienta propuesta para esta investigación (*CECASDEP*) trata de recoger las tendencias actuales del servicio municipal deportivo, considerando todos los procesos que ocurren en su prestación mediante cinco grandes dimensiones con el objetivo de evaluar la percepción de calidad por parte de los/as usuarios/as, apostando así por una actualización de las herramientas existentes.

MÉTODO

Participantes

En el desarrollo de esta investigación se utilizaron dos muestras, una para cada versión de la herramienta *CECASDEP*, administrando la primera versión (*v.1.0*) en los municipios de Vélez-Málaga y Ronda (Málaga) a una muestra compuesta por 110 participantes, de los que 48 (43.6%) son de género femenino y 62 (56.4%) masculino, con edades comprendidas entre los 23 y los 66 años (m = 36.13; dt = 8.41). La segunda versión de la herramienta (*v.2.0*) fue empleada sobre una muestra de 537 participantes, 232 (43.2%) femeninos, 295 (54.9%) masculinos y 10 (1.9%) que no respondieron al género, usuarios/as de los servicios deportivos del municipio de Mijas (Málaga), con una edad media de 32.11 años (dt = 11.11) y un rango de edad que oscila entre los 14 y los 69 años.

Material

Se elaboró el *Cuestionario de Evaluación de la Calidad Percibida en Servicios Deportivos* (*CECASDEP*), cuya primera versión (*v.1.0*) está compuesta por 71

ítems distribuidos en 6 escalas, mientras que la segunda versión (*v.2.0*), se compone de 51 ítems y 5 escalas, incluyéndose en las dos versiones una serie de preguntas de carácter sociodemográfico. Las respuestas se emiten en un continuo del 1 (*nada de acuerdo*) al 5 (*muy de acuerdo*), estando los ítems redactados en la misma dirección, de forma que a mayor acuerdo con el enunciado la puntuación será mayor. El análisis de datos se realizó mediante el paquete estadístico SPSS-v.15.0 (SPSS, 2006).

Procedimiento

Se administró la herramienta *CECASDEP-v.1.0* en una muestra de 110 participantes, adecuada para el cálculo del análisis factorial (Del Barrio y Luque, 2000), realizándose un análisis de fiabilidad y un análisis preliminar para comprobar la pertinencia del análisis factorial. Con los resultados obtenidos y realizándose las modificaciones metodológicamente necesarias de forma justificada, se construyó la segunda versión (*CECASDEP-v.2.0*) utilizándose una muestra de 537 participantes para comprobar sus propiedades psicométricas.

Se siguió el mismo protocolo en la recogida de datos con todos los participantes, que fueron informados sobre cómo debían cumplimentar el cuestionario de forma concreta. La participación fue individual, voluntaria y anónima, y siempre en presencia de los investigadores de este estudio con el fin de solucionar posibles dudas que pudieran generar la compresión de los ítems.

Resultados

Los coeficientes alfa de Cronbach (α) obtenidos para cada escala oscilan entre .711 y .951 para la primera versión (*v.1.0*), y entre .764 y .939 para la segunda (*v.2.0*), siendo satisfactorios en todos los casos mostrando una consistencia interna buena o muy buena (Nunnally, 1976). En ambas versiones, la escala relacionada con el profesor-monitor es la que obtiene un mayor índice de fiabilidad, mientras que el resto de escalas muestran valores ligeramente inferiores en la *v.2.0* (tabla 1).

Tabla 1. Número de ítems, índice de consistencia interna, factores y variancia explicada para cada escala del CECASDEP

Escala		Ítems		Alfa de Cronbach		Factores		Variancia explicada	
v.1.0	v.2.0	v.1.0	v.2.0	v.1.0	v.2.0	v.1.0	v.2.0	v.1.0	v.2.0
1	1	16	10	.817	.764	5	3	65.748%	57.649%
2		9		.711		3		64.942%	
3	2	11	10	.863	.868	3	2	66.319%	57.734%
4	3	14	12	.907	.878	3	2	63.441%	53.770%
5	4	9	9	.886	.800	2	2	67.688%	53.632%
6	5	12	10	.951	.939	1	1	66.360%	64.827%

Para la obtención de los factores, se realizó un análisis de componentes principales con rotación Varimax, que proporciona un reparto más homogéneo de la variancia explicada. De esta forma, de la *v.1.0* resultaron 17 factores, explicando cada escala alrededor del 65% de la variancia, mientras que para la *v.2.0* fueron 10 factores, situándose la variancia explicada en torno al 55% para todas las escalas salvo la relacionada con el profesor-monitor, que mantuvo valores similares en ambas versiones, considerándose en cualquier caso resultados adecuados (tabla 2).

Tabla 2. Criterios de pertinencia de realización del análisis factorial

Escala		Determinante		KMO		Esfericidad de Bartlett	
v.1.0	v.2.0	v.1.0	v.2.0	v.1.0	v.2.0	v.1.0	v.2.0
1	1	.002	.102	.779	.792	χ^2 647.820 p=0.000	χ^2 1211.489 p=0.000
2		.059		.671		χ^2 297.312 p=0.000	
3	2	.007	.021	.835	.884	χ^2 513.371 p=0.000	χ^2 2060.854 p=0.000
4	3	.001	.009	.869	.886	χ^2 795.991 p=0.000	χ^2 2529.288 p=0.000
5	4	.007	.084	.855	.843	χ^2 525.059 p=0.000	χ^2 1317.683 p=0.000
6	5	.000003	.001	.906	.936	χ^2 1318.141 p=0.000	χ^2 4088.212 p=0.000

Para comprobar la pertinencia del análisis factorial exploratorio se calculó el determinante de la matriz de correlaciones, la medida de adecuación muestral de Kaiser-Meyer-Olkin (KMO) y el test de esfericidad de Bartlett. Los análisis estadísticos mostraron un resultado adecuado para el determinante con valores muy bajos en todas las escalas (valores para la *v.1.0* entre .000003 y .059 y valores para la *v.2.0* entre .001 y .102); la medida de adecuación muestral refleja valores muy similares en las dos versiones, siendo buenos o muy buenos siguiendo las indicaciones de Kaiser (1958) y la prueba de esfericidad de Bartlett ha resultado estadísticamente significativa (p=.000) en todas las escalas, por lo que los resultados apoyan la idoneidad para llevar a cabo el análisis al presentar las variables correlaciones altas (Visauta, Martori y Cañas, 2005) (tabla 2).

DISCUSIÓN Y CONCLUSIONES

El presente estudio, que se enmarca dentro de una investigación de mayor alcance, se basa en la necesidad de evaluación de la calidad percibida de los servicios municipales deportivos ofertados en la actualidad. Para ello, la estructura de la herramienta *CECASDEP* sigue algunas de las indicaciones de Luna-Arocas et al. (1998) al tratar de recoger los aspectos relevantes del servicio deportivo desde que el/la usuario/a decide acudir a su realización. El cuestionario cuenta con seis (*v.1.0*) y cinco (*v.2.0*) dimensiones respectivamente, estando así en consonancia con las

expuestas por Hernández Mendo (2001), Calabuig y Crespo (2009) o Nuviala et al. (2008, 2010).

El objetivo del presente trabajo ha sido comprobar la estructura factorial de las dos versiones de la herramienta en dos muestras de usuarios/as. Los resultados muestran significación (p=.000) en todas las escalas para el test de esfericidad de Bartlett e índices buenos o muy buenos para la prueba de adecuación muestral (Kaiser, 1958).

Asimismo, ambas versiones presentan una adecuada consistencia interna obteniendo resultados superiores a .80 prácticamente en todas las escalas, proporcionando así evidencia de la estabilidad de la estructura factorial de la herramienta en diferentes muestras de usuarios/as de servicios municipales deportivos.

Consideramos que estos resultados avalan que la herramienta construida puede ser utilizada para la evaluación de la calidad percibida en servicios deportivos de forma fiable, siendo necesario la realización de un análisis factorial confirmatorio que nos permita comprobar su estructura latente, así como también el cálculo de un adecuado *plan de optimización* mediante el uso de la teoría de la generalizabilidad (Cronbach, Gleser, Nanda y Rajaratnam, 1972), al suponer uno de los aspectos de mayor repercusión actualmente en el ámbito de la gestión deportiva para la optimización de los recursos (Morales Sánchez, 2009).

REFERENCIAS

Afthinos, Y., Theodorakis, N. y Nassis, P. (2005). Customers' expectations of service in Greek fitness centres. Gender, age, type of sport center and motivation differences. *Managing Service Quality, 15*(3), 245-258.

Barrera, R. y Reyes, M. C. (2003). Análisis comparado de las escalas de medición de la calidad de servicio. En *XIII Jornadas Hispano-Lusas de Gestión Científica, 13*, 285-294. Lugo: Universidad de Santiago de Compostela.

Calabuig, F. y Crespo, J. (2009). Uso del método delphi para la elaboración de una medida de calidad percibida de los espectadores de eventos deportivos. *Retos. Nuevas tendencias en Educación Física, Deporte y Recreación, 15*, 18-24.

Calabuig, F., Quintanilla, I. y Mundina, J. (2008). La calidad percibida de los servicios deportivos: diferencias según instalación, género, edad y tipo de usuario en servicios náuticos. *International Journal of Sport Science, 10*(4), 25-43.

Cronbach, L. J., Gleser, G. C., Nanda, H. y Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles.* New York: John Wiley and Sons.

Del Barrio, S. y Luque, T. (2000). Análisis de ecuaciones estructurales. En T. Luque (Coord.), *Técnicas de análisis de datos en investigación de mercados* (489-557). Madrid: Pirámide.

Hernández Mendo, A. (2001). Un cuestionario para evaluar la calidad en programas de actividad física. *Revista de Psicología del Deporte, 10*(2), 179-196.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 183-200.

Ko, Y. J. y Pastore, D. L. (2005). A hierarchical model of service quality for the recreational sport industry. *Sport Marketing Quarterly, 14*(2), 84-97.

Luna-Arocas, R., Mundina, J. y Gómez, A. (1998). La creación de una escala para medir la calidad de servicio y la satisfacción: el Neptuno-1. En J. Martínez del Castillo (Comp.), *Deporte y Calidad de Vida* (279-290). Madrid: Librerías Deportivas Esteban Sanz.

Marmol, A., Orquín, F. J. y Sainz, P. (2010). La infraestructura y el equipamiento, la prescripción del ejercicio y los servicios ofertados como índices de calidad de los centros fitness de Murcia. *Cuadernos de Psicología del Deporte, 10* (Suplem.), 85-91.

Morales Sánchez, V. (2003). *Evaluación psicosocial de la calidad en servicios municipales deportivos: aportaciones desde el análisis de variabilidad.* Málaga: SPICUM.

Morales Sánchez, V. (2009). Evaluación de la calidad en organizaciones deportivas: análisis de generalizabilidad. *Revista de Psicología General y Aplicada, 62*(1-2), 99-109.

Morales Sánchez, V., Blanco, Á. y Hernández Mendo, A. (2004). Optimización de modelos de medida en la evaluación de programas de actividad física. *Metodología de las Ciencias del Comportamiento, Suplemento 2004*, 427-433.

Morales Sánchez, V., Hernández Mendo, A. y Blanco, Á. (2005). Evaluación de la calidad en los programas de actividad física. *Psicothema, 17*(2), 311-317.

Morales Sánchez, V., Hernández Mendo, A. y Blanco, Á. (2009). Evaluación de la calidad en organizaciones deportivas: adaptación del modelo SERVQUAL. *Revista de Psicología del Deporte, 18*(2), 137-150.

Morales Sánchez, V. y Gálvez, P. (2011). La percepción del usuario en la evaluación de la calidad de los servicios municipales deportivos. *Cuadernos de Psicología del Deporte, 11*(Suplem.), 147-154.

Nuviala, A., Tamayo, J. A., Iranzo, J. y Falcón, D. (2008). Creación, diseño, validación y puesta en práctica de un instrumento de medición de la satisfacción de usuarios en organizaciones que prestan servicios deportivos. *Retos. Nuevas tendencias en Educación Física, Deporte y Recreación, 14*, 10-16.

Nuviala, A., Tamayo, J. A., Nuviala, R., González, J. A. y Fernández, A. (2010). Propiedades psicométricas de la escala de valoración de organizaciones deportivas EPOD. *Retos. Nuevas tendencias en Educación Física, Deporte y Recreación, 18*, 82-87.

Parasuraman, A., Zeithaml, V. y Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12-40.

Rial, J., Varela, J., Rial, A. y Real, E. (2010). Modelización y medida de la Calidad Percibida en centros deportivos: la escala QSport-10. *Revista Internacional de Ciencias del Deporte, 18*(6), 57-73.

Salvador, C. M. (2005). La percepción del cliente de los elementos determinantes de la calidad del servicio universitario. Características del servicio y habilidades profesionales. *Papeles del Psicólogo, 90*, abril.

Sanz, I., Redondo, J., Gutiérrez, P. y Cuadrado, G. (2005). La satisfacción en los practicantes de spinning: elaboración de una escala para su medición. *Motricidad: European Journal of Human Movement, 71*(13), 17-36.

Serrano, V., Rial, A., García, Ó. y Hernández Mendo, A. (2010). La evaluación de la calidad percibida del servicio como elemento clave para la gestión de los clubs de golf en España. *Apunts: Educación Física y Deportes, 102*(4), 95-105.

SPSS (2006). *Statistical Package for Social Sciences for Windows.* Version 15. Chicago, IL: SPSS Inc.

Visauta, B., Martori, I. y Cañas, J. C. (2005). *Análisis estadístico con SPSS para Windows.* México: McGraw-Hill.

VALIDACIÓN DE UN CUESTIONARIO PARA EVALUAR LA CALIDAD EN LA ORGANIZACIÓN DEL VOLUNTARIADO DEPORTIVO UNIVERSITARIO

Rosa García González[1], Encarnación Chica Merino[1], Verónica Morales Sánchez[2] y Antonio Hernández Mendo[2]

[1] Universidad de Cádiz

[2] Universidad de Málaga

Correo electrónico: rosa.garcia@magisteriolalinea.com

Resumen

Desde hace ya varios años las universidades, además de su función en el campo de la docencia, la investigación y la formación, decidieron sumarse también al reto de un compromiso social, acercándose a la sociedad para dar respuesta a las demandas de su entorno. De esta forma, colaboran con diferentes instituciones, que también potencian el voluntariado, contribuyendo a través de la participación de los jóvenes universitarios. En este ámbito encontramos entidades deportivas que desde sus programas y acciones organizan eventos deportivos en los que necesitan la participación de voluntarios para su desarrollo. El voluntariado se convierte entonces en una herramienta de acercamiento de la universidad a la sociedad. Por tanto, la gestión, la organización de actividades y programas y la evaluación deben diseñarse bajo una perspectiva distinta que marque el diseño de estrategias propias en el marco del voluntariado. Nuestro estudio se sitúa en el contexto de los programas de voluntariado deportivo, y se centra en la elaboración y validación de un cuestionario cuyo objetivo es evaluar la calidad en la organización de este voluntariado a partir de la satisfacción de los propios voluntarios, que a su vez favorezca la permanencia y fidelidad de los mismos.

Además del voluntariado, existen otras muchas formas de participación ciudadana, basadas en la solidaridad y en el papel activo de la ciudadanía en la construcción de un nuevo modelo social o al menos en un modelo social más participativo y más responsable. Una de estas formas en las que la ciudadanía es co-agente de inclusión social es el aprendizaje servicio.

En este sentido desde la universidad, debemos vincular el grado de satisfacción de voluntariado al grado de satisfacción del aprendizaje servicio para poder establece así las relaciones de su satisfacción general que concibe el sujeto.

Nuestro estudio se centra en la crear una herramienta para valorar el proceso de la organización de los programas de voluntariado evaluando la calidad de la

organización de este voluntariado a partir de la satisfacción de los propios voluntarios, que a su vez favorezca la permanencia y fidelidad de los mismos.

Por ello vemos necesario validar una herramienta (cuestionario) para evaluar dicha organización.

De esta forma, de la colaboración de entidades universitarias con diferentes instituciones, públicas y privadas, que también potencian el voluntariado, surge nuestra investigación cuyos objetivos son:

1. Validar la herramienta diseñada para evaluar la calidad en la organización de programas de voluntariado deportivo.
2. Valorar la repercusión de la satisfacción con el nuevo perfil del voluntario deportivo universitario.

MÉTODO

De la metodología de nuestra investigación decir que:

El material utilizado ha sido *un cuestionario elaborado a partir de otros estudios de voluntariado medioambiental. (Chica Merino, 2009).*

EL CUESTIONARIO, con mejoras de su versión piloto, consta de un total de 46 ítems distribuidos en 6 escalas. Las respuestas se emiten sobre un continuo del 1 al 4, como así hemos querido recoger los datos relativos a la edad, género, área donde han prestado su ayuda, participación en otros voluntariados deportivos y práctica de deporte.

Se ha utilizado para el análisis del mismo el paquete estadístico SPSS v. 17.0

La investigación se compone de un estudio con una muestra de 210 sujetos, sobre la que se realiza un análisis factorial exploratorio y un análisis de fiabilidad.

Previo a llevar a cabo el análisis factorial, se lleva a cabo un examen de la matriz de correlaciones con el objetivo de poner a prueba la pertinencia de dicho análisis (Visauta, 1998), es decir comprobar si sus características son las más adecuadas para realizar un Análisis Factorial.

La distribución de las escalas es la siguiente:

- Escala 1: **ORGANIZACIÓN DEL VOLUNTARIADO** (7 items)
- Escala 2 **LOS/AS RESPONSABLES DE ÁREA** (17 items)
- Escala 3 **TAREAS ESPECÍFICAS DEL ÁREA** (8 items)
- Escala 4 **LOGÍSTICA** (9 items)
- Escala 5 **INSTALACIÓN Y MATERIALES** (4 items)
- Escala 6 **VALORACIÓN PERSONAL** (8 items)

RESULTADOS Y CONCLUSIONES

En la tabla 1 que presentamos a continuación podemos observar los valores que presentan los diferentes indicadores que nos muestran la pertinencia:

Con estos resultados los criterios de pertinencia de realización del análisis factorial se cumplen ya que:

- Los determinantes de la matriz de correlaciones son bajos, siendo el valor más bajo 0,043 para la escala 2, y el más alto 0,414 para la escala 5.
- La medida de adecuación muestral KMO2 (Kaiser-Meyer-Olkin) presenta valores >0.72 y >0,85 para las distintas escalas, por lo que la idea de utilizarse un análisis factorial es aceptable para las escalas 5 y buena para las escalas 1, 2,3,4 y 6.
- El test de esfericidad de Bartlett3 con p<0.05 y valores altos ÷2 para todas las escalas.
- Los valores en la matriz anti-imagen son bajos y en MSA altos.

Tabla 1. Criterios de pertinencia de realización del análisis factorial del cuestionario ECPVE v.2.0

Descriptivos	Escala 1	Escala 2	Escala 3	Escala 4	Escala 5	Escala 6
Determinante	0'286	0,043	0,233	0,023	0,414	0,068
KMO	0,780	0,849	0,807	0,829	0,729	0,850
Test de Bartlett	χ2 257,58 gl 21 sig 0'0	χ2 646,64 gl 45 sig 0'0	χ2 299,60 gl 28 sig 0'0	χ2 771,21 gl 36 sig 0'0	χ2 182,42 gl 6 sig 0'0	*χ2 553,01* *gl 28* *sig 0'0*
MSA	*Min 0,740* *Max 0,808*	*Min 0,790* *Max 0,925*	*Min 0,751* *Max 0,845*	*Min 0 781* *Max 0,881*	*Min 0,690* *Max 0,801*	*Min 0,825* *Max 0,912*

Tabla 2. Análisis Factorial y de la Fiabilidad del Cuestionario (LOS BARRIOS) ECPVE v.2.0

ECPVE v.2.0			
ESCALAS	**Nº factores**	**% Varianza**	**Alfa de Cronbach**
1	3	64,89	,723
2	3	61,53	,844
3	3	61,35	,746
4	3	70,34	,854
5	1	56,08	,730
6	2	58,89	,838

En relación a los resultados del análisis factorial y el análisis de la fiabilidad representado en la tabla vemos que:

- La varianza explicada oscila entre 56,08 para la escala 5 con 1 factor y 70,34 para la escala 4 con 3 factores.

- Respecto a la fiabilidad, el Alpha de Cronbach alcanza valores de 0,723 para la escala 1 y 0,854 para la escala 4, considerándose un valor satisfactorio.
- Las escalas 1, 2 y 3 se refieren a 1 factor, en los resultados se aprecia una estructura factorial simple que explica un porcentaje de varianza del 64,89 para la escala 1 y 61,53; 61,35 para la 2 y 3 respectivamente.
- La escala 6 se divide en dos factores, en los resultados se aprecia una estructura factorial simple que explica un porcentaje de varianza del 58,89%
- La escala 5 se aprecia una estructura factorial simple que explica un porcentaje de varianza del 68,536 %,
- En la matriz de componentes rotados, los índices que representan las saturaciones de los dos factores oscilan entre 0,585 y 0,905 (escala 1). Consideramos que son representativos al tener un valor >0,40.
- En la matriz de componentes rotados, los índices que representan las saturaciones de los dos factores oscilan entre 0,517 y 0,833 (escala 3). Consideramos que son representativos al tener un valor >0,40.
- Por lo que respecta a la fiabilidad, ésta ha sido estimada a través del Alpha de Cronbach y estos índices de fiabilidad resultan satisfactorio ya que se sitúan por encima del 0,72 llegando a alcanzar su punto más alto la escala 4 0,854.

Queremos concluir esta investigación confirmando que la herramienta es válida para evaluar la calidad de los programas de voluntariado deportivo aunque no está exento de múltiples modificaciones fruto de un estudio mucho más amplio que nos ocupará el futuro.

REFERENCIAS

Chica Merino, E. (2009). Construcción de una herramienta para evaluar la calidad de los programas de voluntariado ambiental [Tesis Doctoral].

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles.* New York: John Wiley and Sons.

Hernández Mendo, A. (2001). Un cuestionario para evaluar la calidad en programas de actividad física. *Revista de Psicología del Deporte, 10,* 179-196.

Morales Sánchez, V., Blanco Villaseñor, A., & Hernández Mendo, A. (2004). Optimización de modelos de medida en la evaluación de programas de actividad física. *Metodología de las Ciencias del Comportamiento. Vol. Especial.* 437-443.

Nunnally, J.C. (1976). *Psychometric theory.* New York: McGraw-Hill.

Observatorio del Tercer Sector (2009). *Manual de Gestión del voluntariado*. Obra Social «Fundación La Caixa».

Padilla, J.L., González, A y Pérez, C. (1998). Elaboración del cuestionario. En A.J. Rojas, J.S. Fernández y C. Pérez, Investigar mediante encuestas. Madrid: Síntesis.

Schlotzhauer, S.D. & Littell, R.C. (1997). *SAS System for Elementary Statistical Analysis*. Cary, NC: SAS Institute Inc.

VV.AA. (2009). Manual de Buenas Prácticas en la Gestión de la Proyección Social y el Voluntariado Universitario. Agencia Andaluza del Voluntariado.

Vecina Jiménez, M. L., Chacón Fuertes, F., & Sueiro Abad M. J. (2009). Satisfacción en el voluntariado: estructura interna y relación con la permanencia en las organizaciones. *Psicothema,* 21(1), 112-117.

Visauta, B., Martori, I., & Cañas, J.C. (2005). *Análisis estadístico con SPSS para Windows.* México: McGraw-Hill.

LA ESCALA QGOLF-9 PARA LA GESTIÓN DE CLUBES DE GOLF CON CAMPOS DE 9 HOYOS. UN ESTUDIO PREVIO

Virginia Serrano Gómez[1], Antonio Rial Boubeta[2], Óscar García García[3] y Vicente Gambau i Pinasa[1]

[1] Universidad de A Coruña
[2] Universidad de Santiago de Compostela
[3] Universidad de Vigo
Correo electrónico: virginia.serrano@udc.es

Resumen

En el contexto de un estudio más amplio se desarrolla la escala QGOLF-9, la misma permite medir la calidad percibida del servicio prestado en clubes de golf con campos de 9 hoyos. Recoge tres dimensiones generales del servicio: (1) organización y profesionalidad, (2) instalaciones y servicios complementarios y, (3) campo y zona de juego. Dada su brevedad y sencillez (15 ítems tipo Likert) constituye un instrumento muy manejable, que además presenta una alta consistencia interna, tanto a nivel global (α de Cronbach=0.91) como para cada una de sus subescalas o dimensiones, así como una capacidad para explicar la satisfacción de los usuarios/as realmente interesante. La muestra de este estudio previo estuvo formada por 338 individuos (263 hombres y 75 mujeres), socios-usuarios/as de 5 clubes de golf sociales-mixtos ubicados en Galicia, con edades comprendidas entre los 14 y 75 años (X = 47.22; S_X = 13.97). Tras un primer análisis descriptivo, se realizó un Análisis Factorial Confirmatorio y un análisis *path* para estudiar la validez de constructo y de criterio de la escala, respectivamente. Por último se realizó un Análisis IPA (*Importance Performance Analysis*) que posibilita un diagnóstico estratégico del funcionamiento de este tipo de clubes y se traduce en recomendaciones concretas para la gestión.

Tanto los juicios de calidad, como la satisfacción del usuario constituyen elementos clave en la estrategia de las nuevas empresas u organizaciones deportivas, lo que ha llevado a sus responsables a establecer planes de calidad y definir acciones específicas que les permitan alcanzar una mayor satisfacción de sus clientes, optimizar los recursos disponibles, y favorecer la mejora continua (Luna, Mundina y Carrión, 1998; Morales Sánchez, 2003). Bajo este nuevo enfoque los clientes y/o usuarios se han convertido en los principales protagonistas del servicio, y los esfuerzos de la organización se centran en disponer de un mejor conocimiento de éste. En esta línea, la necesidad de disponer de herramientas que ayuden en la gestión a través de una adecuada evaluación de la *calidad percibida*, ha hecho que

se hayan desarrollado instrumentos en campos muy diferentes incluido el deportivo, sin embargo en el contexto específico del golf no se encuentran herramientas específicas que evalúen la calidad percibida desde el punto de vista de sus usuarios de forma válida, fiable, y manejable en términos de gestión (Serrano, Rial, García, y Hernández-Mendo, 2010). Para ello el objeto de estudio será elaborar una escala breve y de fácil aplicación que permita incorporar una medida de calidad percibida a la gestión de los clubes de golf y más concretamente con campos de 9 hoyos, como estudio previo a otro de mayor dimensión; relacionar los elementos de la escala y la satisfacción o *performance*; y utilizar el Análisis de Importancia Valoración como técnica para realizar un diagnóstico del desempeño de estos clubes.

MÉTODO

Participantes

Para dar cuenta de los objetivos se utilizó una encuesta entre usuarios de clubes de golf mixtos con campos de 9 hoyos en Galicia. La muestra estaba formada por 338 participantes, 263 hombres y 75 mujeres con edades entre 14 y 85 años (X = 47,22; S_X= 13,97).

Instrumentos y medidas

El instrumento para recoger la información fue un cuestionario *ad hoc*. Se utilizó una adaptación de la escala de calidad percibida para clubes de golf de Serrano, Rial, García y Hernández-Mendo (2010). La misma, estaba formada inicialmente por 25 ítems y 4 dimensiones obtenidas a partir de un Análisis Factorial Confirmatorio. Con una consistencia interna elevada (α=0.92) y una alta capacidad para explicar la satisfacción del usuario (R^2=0.62). Además se utilizó la adaptación del Análisis de Importancia Valoración (IPA) de Abalo, Varela y Rial (2006) donde se representan las discrepancias entre la valoración y la importancia otorgadas por los usuarios.

Procedimiento

El procedimiento para la recogida de datos fue la entrevista personal. Los participantes se eligieron de forma accidental y se entrevistaron personalmente por personal formado para ello, durante distintos días y diferentes horas, con una duración aproximada de 10-12 minutos por usuario entrevistado.

Los datos fueron analizados con el paquete estadístico PASW Statistics 18 y el AMOS 16.

RESULTADOS

En la tabla 1 se recoge las medias y desviaciones típicas de los 25 ítems iniciales, así como los valores del alfa de Cronbach e índice de homogeneidad de cada elemento.

Tabla 1. Descriptivos de la escala de 25 elementos de calidad percibida del servicio de golf

	Dimensión/Ítem. ELEMENTOS DEL SERVICIO	Media	Desv. típica	IHC	α de Cronbach
D.1/Q1	Profesionalidad Gerencia	3,77	1,38	0,69	0,93
D.1/Q2	Profesionalidad Recepción	3,90	1,12	0,65	0,93
D.1/Q3	Profesionalidad Greenkeeper	4,08	1,15	0,58	0,93
D.1/Q4	Profesionalidad Master Caddie	4,10	1,20	0,64	0,93
D.1/Q5	Profesionalidad Profesores/as Golf	4,49	1,07	0,42	0,93
D.2/Q6	Implicación de la Directiva	3,75	1,33	0,72	0,93
D.2/Q7	Organización y gestión de los recursos del club	3,80	1,36	0,73	0,93
D.2/Q8	Gestión de la información/comunicación	3,64	1,33	0,75	0,93
D.2/Q9	Gestión de las reclamaciones y sugerencias ,rápida y eficaz	3,89	1,62	0,63	0,93
D.2/Q10	Seguridad y prevención de riesgos (medidas caso de emergencia)	4,18	1,35	0,59	0,93
D.2/Q11	Gestión medioambiental	4,32	1,20	0,54	0,93
D.2/Q12	Organización de torneos en el club	3,95	1,19	0,71	0,93
D.2/Q13	Correspondencias con otros clubes	4,10	1,48	0,56	0,93
D.2/Q14	Trato y atención	4,19	0,99	0,67	0,93
D.3/Q15	Limpieza e higiene general	3,98	0,91	0,64	0,93
D.3/Q16	Estado de las instalaciones del club	3,92	0,90	0,63	0,93
D.3/Q17	Estado del mobiliario-material-equipamientos del club	3,92	0,97	0,63	0,93
D.3/Q18	Salón social-Casa club	3,96	1,02	0,54	0,93
D.3/Q19	Vestuarios	3,83	1,05	0,59	0,93
D.3/Q20	Academia-Escuela de golf	4,15	1,21	0,52	0,93
D.4/Q21	Estado del campo	4,20	0,93	0,47	0,93
D.4/Q22	Control de juego-Cumplimiento del reglamento en el campo	3,53	1,29	0,63	0,93
D.4/Q23	Diseño y recorrido del campo	4,13	0,92	0,41	0,93
D.4/Q24	Seguridad del campo	3,71	1,10	0,55	0,93
D.4/Q25	Zona de prácticas	3,67	1,15	0,48	0,93

La válidez del contructo se realiza a partir de un análisis factorial confirmatorio (AMOS 16). Los primeros resultados resultan estadísticamente significativos, sin embargo el ajuste global es muy pobre tal y como reflejan los distintos indicadores de bondad de ajuste considerados en la tabla 2, llevando a una reespecificación del modelo.

Tabla 2. Indicadores de bondad de ajuste del modelo inicial

	χ^2	gl	p	χ^2/gl			
Modelo inicial	922.66	269	<0.001	3.43			
	NFI	**CFI**	**GFI**	**AGFI**	**TLI**	**RMR**	**RMSEA**
Modelo inicial	0.80	0.85	0.81	0.77	0.83	0.09	0.09

Fruto de esta reespecificación, hubo una reestructuración del modelo que estuvo compuesto por sólo 3 factores. Se refundió las dos primeras dimensiones pres-

cindiendo de algunos ítems con una correlación muy alta y manteniendo las otras dos dimensiones con algunas reducciones. De esta forma el ajuste pasa a ser excelente (tabla 3).

Tabla 3. Indicadores de bondad de ajuste del modelo final

	χ^2	gl	p	χ^2/gl			
Modelo inicial	155.77	84	<0.001	1.87			
	NFI	**CFI**	**GFI**	**AGFI**	**TLI**	**RMR**	**RMSEA**
Modelo inicial	0.94	0.97	0.94	0.92	0.97	0.05	0.05

Así queda finalmente una escala de sólo 15 ítems, bastante manejable y que tiene una alta consistencia interna tanto a nivel global como para cada una de sus dimensiones o subescalas (tabla 4).

Tabla 4. Consistencia interna (a de Cronbach)

PROFESIONALIDAD ORGANIZACIÓN GESTIÓN (7)	**INSTALACIONES (4)**	**CAMPO (4)**	**GLOBAL (15)**
0.89	0.84	0.79	**0.91**

Por otro lado, para relacionar los elementos de la escala y la satisfacción o *performance*, se consideran tres elementos: *valoración global, expectativas y satisfacción* que tienen una gran capacidad predictora, 0´62 de la varianza explicada, y que además tiene una relación importante con los resultados en términos actitudinales (Figura 1).

Con estos datos, y una vez que se comprueba que puede tratarse de una escala adecuada, el diagnóstico de la calidad percibida responde a una puntuación global media de 3,96 puntos sobre 5, las expectativas con respecto al club para la mayoría de los usuarios/as responden tal y *como esperaban*, y la satisfacción general de los mismos es de 7´02 sobre 10 puntos. De esta forma se podría decir que los usuarios/as están en general satisfechos con la prestación del servicio de golf en sus clubes.

Sin embargo, aunque los resultados obtenidos del diagnóstico de la calidad percibida son *a priori* positivos, al introducir las medias de la importancia y su diferencia con la valoración, casi todos los elementos tienen discrepancias negativas. Por tanto el desempeño está por debajo de lo que se espera de él.

Los resultados de las discrepancias se representan en la Figura 2, donde se observa como el elemento *profesionalidad de los profesores de golf,* es el único elemento que la gerencia puede considerar como aceptable para los usuarios, mientras que los que necesitan mayor atención y en consecuencia pueden estar creando las principales fuentes de insatisfacción entre los usuarios son: *implicación de la directiva, control del juego-cumplimiento del reglamento en el campo,* y *gestión de la información-comunicación.*

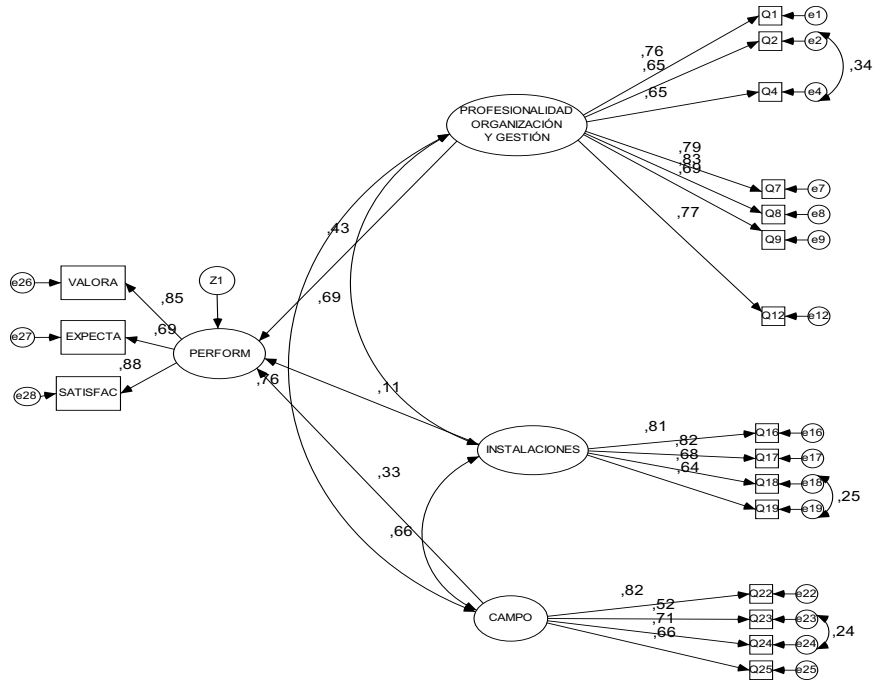


Figura 1. Validez de criterio. Relaciones Calidad Percibida- PERFORMANCE

Profesionalidad Profesores/as Golf
Seguridad y prevención de riesgos (medidas caso de emergencia)
Gestión medioambiental
Trato y atención
Profesionalidad Greenkeeper
Correspondencias con otros clubes
Academia-Escuela de golf
Diseño y recorrido del campo
Gestión de las reclamaciones y sugerencias ,rápida y eficaz
Salón social-Casa club
Profesionalidad Master Caddie
Estado del mobiliario-material-equipamientos del club
Estado del campo
Profesionalidad Recepción
Vestuarios
Profesionalidad Gerencia
Estado de las instalaciones
Organización de torneos en el club
Limpieza e higiene general
Zona de prácticas
Seguridad del campo
Organización y gestión de los recursos del club
Implicación de la Directiva
Control de juego-Cumplimiento del reglamento en el campo
Gestión de la información/comunicación

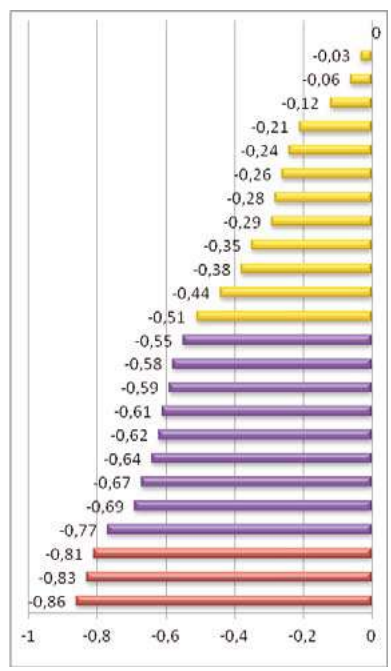


Figura 2. Discrepancias de la Importancia Valoración

Finalmente en la Figura 3, se representa el análisis de la importancia valoración, utilizando para ello la versión de Abalo, Varela y Rial (2006). Se decide por su proximidad a la diagonal que los elementos (1-2-3-4): *profesores, prevención de riesgos, gestión medioambiental* y *trato - atención prestadas,* generan satisfacción acorde con la importancia. Sin embargo los elementos (21-22-23-24) que se separan claramente de la diagonal, son los que van a presentar más insatisfacción y por consiguiente son elementos que requieren mayor prioridad en la mejora de su gestión. Los mismos están representados también en la misma gráfica (3) y corresponden a los siguientes elementos: *gestión de los recursos, implicación de la directiva, control del juego, y gestión de la información –comunicación*.

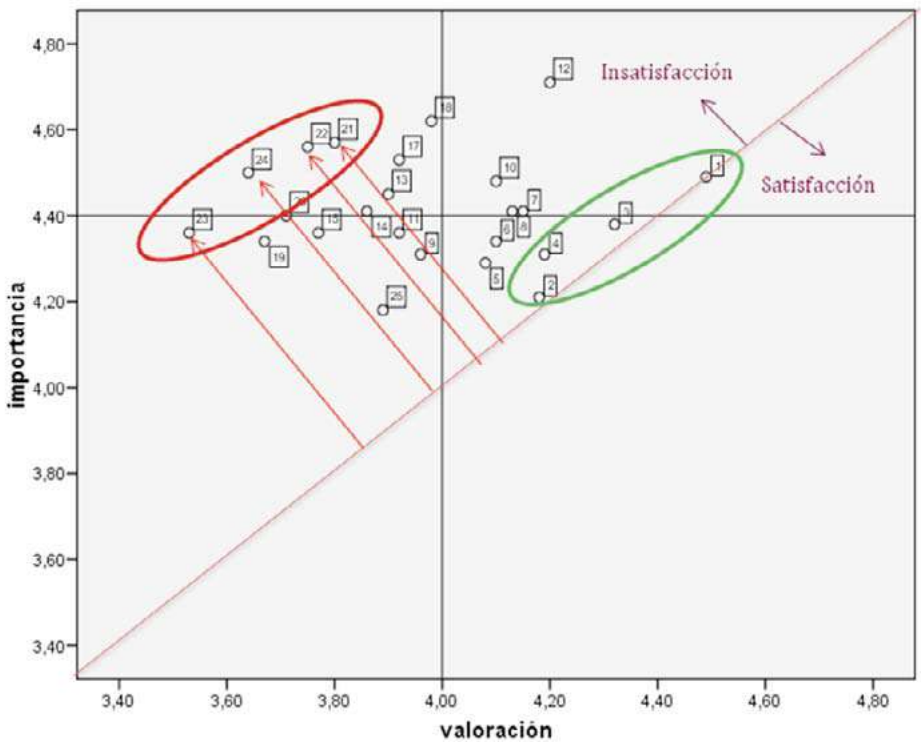


Figura 3. Representación Importancia Valoración del servicio de golf a partir de las discrepancias

Además es destacable, la ausencia de discrepancias positivas o puntos en el área de la satisfacción, lo que indica que existen elementos del servicio que están sobre la diagonal, a la altura de lo que se espera, mientras que el *performance* resultante se encuentra por debajo. De ello se deduce que los usuarios de golf son bastante exigentes, y otorgan mucha importancia a todos los elementos del servicio. Por tanto los responsables de estas instalaciones no deberán descuidar ningún elemento, y especialmente los más alejados de la diagonal.

DISCUSIÓN Y CONCLUSIONES

La escala desarrollada daría cuenta de 3 grandes dimensiones de la calidad percibida: *1) Profesionalidad, Organización y gestión, 2) Instalaciones y servicios complementarios, 3) Campo-Zona de juego.* Tanto la escala global como las tres dimensiones que la componen presentan una consistencia muy elevada 0´91. La misma permite explicar el 62% del desempeño, *performance,* o en términos actitudinales la satisfacción de los usuarios/as con el servicio.

Por otro lado, se ha utilizado como herramienta de diagnóstico el Análisis de Importancia Valoración (*Importance Performance Analysis: IPA*) concretamente la adaptación de Ábalo, Varela y Rial (2006) reformulación de la gráfica original de Martilla y James (1977). Esta técnica ha sido aplicada en distintos ámbitos, incluido el deportivo (Rial, Rial, Varela y Real, 2008; Tarrant & Smith, 2002; Yildiz, 2011), la misma se presenta como una herramienta muy útil que permite identificar fácilmente las áreas fuertes y deficitarias del servicio, estableciendo un rápido diagnóstico del funcionamiento de las mismas y orientando cuales necesitan mayor o menor atención. Sin embargo no se encuentra en la literatura referencias de esta aplicación circunscritas al ámbito específico del golf, por lo que la aplicación del IPA en este contexto, posibilita un diagnóstico estratégico del funcionamiento y orientación de los clubes de golf hacia la mejora continua. En este contexto, los datos de este estudio revelan que los elementos más relevantes del servicio para los usuarios de golf son: *1) Estado del campo 2) Limpieza e higiene 3) Organización y gestión de los recursos del club.*

Además, la existencia de tantas discrepancias negativas entre la importancia y valoración de los elementos, parece indicar que el nivel de exigencia de los usuarios es muy alto, lo que supondrá un reto para los gestores y clubes de golf modernos. Los elementos más débiles y por tanto con una mayor discrepancia negativa son: *gestión de la información, control del juego-cumplimiento del reglamento en el campo, e implicación de la directiva,* lo que permite orientar a efectos de gestión hacia donde deben dirigirse los esfuerzos principalmente.

Finalmente con el objeto de realizar un diagnóstico estratégico del funcionamiento de los clubes de golf desde el punto de vista del socio o usuario, identificando áreas prioritarias para la gestión, se distingue de las cuatro áreas estratégicas de Ábalo, Varela, y Rial (2006) las siguientes dos: Área I. Concentrarse aquí: *Gestión recursos, Implicación directiva, Control del juego, Información/comunicación.* Área II. Mantener el buen trabajo: *Profesores. Prevención de riesgos. Gestión medioambiental. Trato.* Como puede observarse no es posible hablar en este estudio de elementos de baja prioridad, ni de un posible derroche de recursos, centrándose la mayoría de los puntos en el área *concentrarse aquí*. En trabajos como el de Oh (2001) o incluso el de Ábalo, Varela, y Rial (2006) en gestión de servicios sanitarios, también surge la problemática de que la mayoría de los atributos recaen en un mismo cuadrante, Hollenhorst, Olson & Fortney (1992) ya afirmaban que esta era una limitación habitual en el análisis de importancia valoración.

Para ello Ábalo, Varela y Rial, (2006) optó por situar los ejes en la media de la puntuaciones de cada dimensión posibilitando distribuir los atributos positivos en función de sus puntuaciones relativas de importancia y valoración, reuniendo las ventajas de la representación clásica por cuadrantes y la representación por discrepancias. Siguiendo estas pautas y situando el eje en la media de las dimensiones se reparten las puntuaciones en otros cuadrantes, sin embargo los resultados obtenidos siguen situándose por encima de la diagonal y por tanto parece que los usuarios de golf no están totalmente satisfechos con los elementos del servicio evaluados. Lo que lleva a seguir trabajando en esta línea.

REFERENCIAS

Abalo, J., Varela, J., y Rial, A. (2006). El análisis de Importancia-Valoración aplicado a la gestión de servicios. *Psicothema, 18,* (4), 730-737.

Hollenhorst, S., Olson, D., Fortney, R. (1992). Use of Importance-Performance Analysis to Evaluate State Park Cabins: The Case of the West Virginia State Park System. *Journal of Park and Recreation Administration*, 10, 1-11.

Luna, R., Mundina, J. y Carrión, C. (1998). La satisfacción del consumidor en un centro deportivo. En Martínez del Castillo, J. (comp.) *Deporte y Calidad de Vida*. (Pp. 299-305). Madrid: Librerías Deportivas Esteban Sanz.

Martilla J.A. & James, J.C. (1977). Importance- Performance Analysis. *Journal of Marketing, 41*, 77-79.

Morales Sánchez, V. (2003). *Evaluación psicosocial de la calidad en los servicios municipales deportivos: aportaciones desde el análisis de variabilidad*. Tesis doctoral. Universidad de Málaga.

Oh, H. (2001). Revisiting importance-performance analysis. *Tourism Management, 22*, 617-627.

Rial, A., Rial, J., Varela, J. y Real, E. (2008). An application of Importance-Permormance Analysis (IPA) to the management of Sports Centres. *Managing Leisure, 13*, 179-188.

Serrano, V., Rial, A., García, O., Hernández, A. (2010). La evaluación de la calidad percibida del servicio como elemento clave para la gestión de los clubs de golf en España. *Apunts. Educación Física y Deportes.102* (4), 96-106.

Tarrant, M. A. and Smith, E. K. (2002). The use of a modified importance-performance framework to examine visitor satisfaction with attributes of outdoor recreation settings, Managing Leisure, 7, 69–82.

Yildiz, S.M. (2011). An importance-performance analysis of fitness center service quality: Empirical results from fitness centers in Turkey. *African Journal of Business Management, 5*(16), 7031-7041.

INVENTARIO DE CALIDAD EN LOS CENTROS DE ATENCIÓN INFANTIL TEMPRANA: ANÁLISIS FACTORIAL EXPLORATORIO

Rita P. Romero Galisteo[1,2], Verónica Morales Sanchez[2] y Eduardo Sánchez Guerrero[2]

[1] CAIT del Excmo. Ayto. de Antequera (Málaga)

[2] Universidad de Málaga

Correo electrónico: rpromero@uma.es

Resumen

La *calidad* es uno de los elementos estratégicos en que se fundamenta la mejora de los sistemas organizativos modernos. En toda clase de organizaciones vinculadas al sector servicios, cuando se analiza la *calidad* se obtiene que no esta exenta de problemas dada la dificultad que plantea la definición y medida de la misma. La *atención temprana* (AT) es un ámbito de trabajo reciente, por lo que estamos asistiendo a modificaciones importantes en cuanto a metas, modelos, métodos, sin olvidar la propia gestión de la *calidad de servicio*. Hemos pasado de un modelo de intervención centrado en el niño a uno que tiene en cuenta además de a éste, a su familia y al entorno. Actualmente, cualquier acción sobre las personas no se entiende sin tener una medida de calidad del proceso de intervención y de la calidad percibida por los usuarios. Pese a los sistemas de acreditación, los servicios de AT están necesitados de una herramienta pragmática que permita conocer la opinión de sus clientes. Una vez determinada esta necesidad, corresponde tratar de establecer cuáles pueden ser los contenidos mínimos que se necesitan para evaluar la calidad en los *centros de atención temprana*, objeto de esta comunicación. Uno de los modelos más comunes para analizar ciertos constructos es el análisis factorial exploratorio (AFE). Si no se posee una concepción previa de la estructura del constructo, como es nuestro caso, el uso del AFE es adecuado para purificar los datos y contribuye a la clarificación conceptual y desarrollo de mejores instrumentos de medida.

Cuando hablamos de *atención temprana* se abre un universo de ideas y de acciones. La AT es un concepto amplio, cargado de matices que incluyen múltiples acciones sobre distintos ámbitos donde el niño se desenvuelve durante las primeras etapas del desarrollo.

El concepto de AT ha ido evolucionando a lo largo de las últimas décadas, pasando de un modelo tradicional, imperante en los años 70-80, hasta el modelo actual basado en una concepción biopsicosocial. El modelo tradicional, también lla-

mado, *estimulación precoz*, planteaba una intervención dirigida fundamentalmente al niño. Estaba basado en el entrenamiento sensoriomotor y utilizaba criterios conductuales para enseñar al niño habilidades concretas. El objetivo era potenciar al máximo las capacidades del niño y para ello se ponían en marcha actividades sistemáticas que pretendían mejorar los niveles madurativos en las distintas áreas.

Una de las definiciones, actualmente en vigor, concibe la *atención temprana* como el conjunto de intervenciones, dirigidas a la población infantil de cero a seis años, a la familia y al entorno, que tiene por objetivo dar respuesta lo más pronto posible a las necesidades transitorias o permanentes que presentan los niños con trastornos en su desarrollo o que tienen el riesgo de padecerlos. Estas intervenciones, que deben considerar la globalidad del niño, han de ser planificadas por un equipo de profesionales de orientación *interdisciplinar* o *transdisciplinar* (Grupo de Atención Temprana, 2000).

Estos profesionales, pioneros de una nueva disciplina, sentaron las bases que han hecho posible que lleguemos hasta nuestros días. Su labor nos ha permitido observar, evaluar y aprender, marcando la línea de actuación actual, es decir, la intervención debe considerarse no sólo con el niño, sino también con su entorno familiar y escolar (Perpiñán, 2009).

En las últimas décadas el interés por la calidad no ha parado de crecer. A la presión por la competitividad, que demanda modelos que ayuden a comprender el comportamiento y las evaluaciones que realizan los usuarios y consumidores, se une la emergencia de una nueva gestión de los servicios públicos que intenta compatibilizar sus objetivos sociales con una atención al usuario de mayor calidad. Asimismo, otras organizaciones que se encargan de la defensa y el asesoramiento de los usuarios y consumidores requieren de conocimientos de esta índole para realizar de manera adecuada sus funciones.

Los responsables de las empresas y las organizaciones de servicios no deben quedar indiferentes ante los cambios que, en este sentido, se están produciendo en nuestra sociedad, ya que el éxito de su gestión depende, en parte, de conocer las necesidades y satisfacción de las mismas que tienen sus clientes (Martínez-Tur, Peiró y Ramos, 2001).

En este sentido y dado que la AT no es ajena a la creciente preocupación por la calidad que impregna a la sociedad actual, se plantea el principal objetivo de nuestra investigación, diseñar una herramienta de medida que evalúe la calidad percibida por los principales usuarios indirectos de los CAIT, los familiares de los niños/as atendidos en este tipo de centros (Romero Galisteo y Morales Sánchez, 2010a).

Asimismo, la literatura sobre análisis factorial afirma que el objetivo del mismo es identificar las estructuras principales o dimensiones que subyacen sobre los factores originales y reducir el número de factores, con la pérdida mínima de información. Los modelos más comunes para analizar ciertos constructos son el análisis factorial exploratorio (AFE), que se caracteriza porque no se conocen *a priori* el número de factores en la aplicación empírica donde se determina este número. Por

tanto, si no poseemos una concepción previa de la estructura del constructo, como es nuestro caso, el uso del AFE es adecuado para purificar los datos y contribuye a la clarificación conceptual y desarrollo de mejores instrumentos de medida, objetivo de nuestra investigación (Tomás Miguel, 1993).

MÉTODO

Participantes

Los participantes en el estudio se dividen en 2 bloques: 37 *encuestadores*, que ayudaron en la recogida de información, y 672 *usuarios*, en este caso los padres y madres de los niños-as que reciben tratamiento de atención temprana en catorce de los dieciséis CAIT existentes en la provincia de Málaga.

Material

Para el estudio global se utilizaron:

- Paquete estadístico SPSS para Windows v. 15.0.: Para el tratamiento de los datos recogidos de la muestra de los dos estudios, tanto el pilotaje de la herramienta (con una muestra de 102 usuarios) como el estudio final con una muestra de 672 usuarios.
- Cuestionario: Utilizamos par este estudio la segunda versión del cuestionario elaborado para tal fin y al que hemos denominado Inventario de Calidad en los Centros de Atención Infantil Temprana (ICCAIT- v. 2.0), instrumento que evalúa la calidad en los centros de atención infantil temprana (Romero Galisteo y Morales Sánchez, 2010,a,b,c).

Procedimiento

Para alcanzar uno de los objetivos de nuestra investigación, ya que esta comunicación es parte de un estudio más amplio, el desarrollo de la misma se llevó a cabo en dos fases bien diferenciadas. En primer lugar se realizó un estudio piloto con la primera versión de nuestro cuestionario. A continuación realizamos un estudio con la segunda versión del mismo y una muestra de participantes más amplia, fruto del cual exponemos a continuación los resultados obtenidos.

El *ICCAIT-v.2.0* cuenta con 48 ítems distribuidos en 6 escalas que recogen información relativa a las instalaciones, a las salas de tratamiento y material, a la atención al usuario, al personal especializado, a la información general y a la información técnica. Esta segunda versión de la herramienta surge de la aplicación de los resultados obtenidos tanto del análisis factorial exploratorio como en el análisis de fiabilidad efectuados en el estudio piloto con la primera versión.

En esta versión además, se ampliaron los datos sociodemográficos, solicitando ahora información sobre el parentesco de la persona que responde al cuestionario con el usuario/a, así como sobre si ha acudido anteriormente a otro CAIT o cuánto tiempo lleva acudiendo al actual. Para terminar, ofrecemos la posibilidad de realizar observaciones y/o sugerencias

Mediante el AFE se evaluaron las propiedades psicométricas de nuestro cuestionario, obteniendo la estructura subyacente de nuestra herramienta así como las interrelaciones entre variables. De la misma manera, se determinó la validez del constructo analizado y se optimizó la herramienta de evaluación propuesta. Asimismo se indentificó un número determinado de factores para lo cual necesitamos someter nuestros datos a un análisis de fiabilidad.

RESULTADOS

En primer lugar analizamos la pertinencia de realizar un AFE. Los valores obtenidos en el determinante de la matriz, la medida de adecuación muestral Kaiser-Meyer-Olkin (KMO) y el test de esfericidad de Bartlett nos indican que se cumplen los criterios de pertinencia de realización del análisis factorial (tabla 1).

Tabla 1. Resultados del AFE del cuestionario ICCAIT-v.2.0

Descriptivos	Determinante	KMO	Test esfericidad de Bartlett
Escala 1 (ítems del 1 al 13)	.022	.868	χ^2: 2506.401 gl: 78 sig: .000
Escala 2 (ítems del 14 al 22)	.013	.888	χ^2: 2854.960 gl: 36 sig: .000
Escala 3 (ítems del 23 al 29)	.218	.790	χ^2: 1005.171 gl: 21 sig: .000
Escala 4 (ítems del 30al 37)	.048	.886	χ^2: 2013.523 gl: 28 sig: .000
Escala 5 (ítems del 38 al 42)	.868	.528	χ^2: 93.864 gl: 10 sig: .000
Escala 6 (ítems del 43 al 48)	.062	.874	χ^2: 1857.771 gl: 15 sig: .000

Una vez realizados diferentes análisis factoriales exploratorios así como de fiabilidad de las seis escalas que componen el *ICCAIT-v.2.0* con objeto de optimizar nuestra herramienta de evaluación y analizar la consistencia interna de la mis-

ma observamos que tanto los valores del KMO como del % de varianza explicado mejoraban significativamente respecto a los obtenidos inicialmente, como se observa en la tabla 2.

Tabla 2. Resultados comparativos entre el ICCAIT-v.2.0 y su optimización

Escala	KMO inicial	% Varianza inicial	KMO final	% Varianza final
1	.868	54.06	.868	57.875
2	.888	63.085	.885	69.850
3	.790	41.673	.790	41.673
4	.886	50.011	.884	56.494
5	.528	49.975	.528	49.975
6	.874	61.417	.874	61.417

Observando las *matrices de componentes rotados* de cada escala para ver en cuántos factores satura cada una, obtenemos que el número de factores en los que saturan los 48 ítems del ICCAIT-v.2.0 es diez, a los que hemos denominado:

- sala de espera,
- condiciones ambientales del centro,
- ubicación del centro,
- material,
- salas de tratamiento,
- atención al usuario,
- cualificación, coordinación y cercanía del personal al usuario,
- sugerencias, reclamaciones y quejas,
- derivación al centro e
- información técnica.

CONCLUSIÓN

El objetivo de nuestra comunicación era mostrar los resultados del estudio de las propiedades psicométricas del cuestionario Inventario de Calidad en los Centros de Atención Infantil Temprana mediante el AFE. Los resultados obtenidos y expuestos anteriormente, muestran que la herramienta propuesta para evaluar la calidad en los CAIT, cumple con las condiciones de fiabilidad y validez satisfactorias. Se observa además una estructura factorial adecuada que determina la pertinencia de realización de dicho análisis.

Como hemos comentado, fue necesario a su vez, realizar un análisis de fiabilidad utilizado para optimizar la herramienta de medida propuesta. La modificación con respecto al primer AFE consistió en eliminar los ítems que tenían valores más

altos en la columna *alpha de Cronbach si se elimina un elemento* de la tabla *estadísticos total-elemento* del *análisis de fiabilidad.* De esta manera además, se mejoraba considerablemente la fiabilidad de las escalas que componen el cuestionario.

REFERENCIAS

Grupo de Atención Temprana (G.A.T.) (2000). *Libro Blanco de la Atención Temprana*. Real Patronato de Prevención y de Atención a Personas con Minusvalía. Ministerio de Trabajo y Asuntos Sociales. Madrid: Autor

Martínez-Tur, V., Peiró, J. M. y Ramos, J. (2001). *Calidad de servicio y satisfacción del cliente.* Madrid: Síntesis Psicología.

Perpiñán, S. (2009). *Atención Temprana y Familia. Cómo intervenir creando entornos competentes*. Madrid: Narcea.

Romero Galisteo, R. P. y Morales Sánchez, V. (2010a). Calidad y atención temprana: breve revisión teórica. *Lecturas: Educación Física y Deportes. Revista Digital, 145.* Junio, 2010.

Romero Galisteo, R. P. y Morales Sánchez, V. (2010b). Atención temprana y calidad de servicio. Propuesta de una herramienta de medida. *Lecturas: Educación Física y Deportes. Revista Digital, 145.* Junio, 2010.

Romero Galisteo, R.P. y Morales Sánchez, V. (2010c). *Estimación de diseños para evaluar la calidad en los centros de atención infantil temprana (CAIT).* Libro de actas del XI Congreso de Metodología de las Ciencias Sociales y de la Salud, 149-153. Málaga: 15-18 septiembre.

Tomás Miguel, J. M. (1993). *El uso de los modelos de ecuaciones estructurales y del análisis factorial confirmatorio en el análisis psicométrico de cuestionarios: una batería de seguridad laboral.* Universidad de Valencia: Tesis doctoral.

ANÁLISIS FACTORIAL CONFIRMATORIO: CUESTIONARIO PARA LA EVALUACIÓN DE LA CALIDAD EN PROGRAMAS DE VOLUNTARIADO AMBIENTAL

Encarnación Chica[1], Verónica Morales[2] y Antonio Hernández[2]
[1] Universidad de Cádiz
[2] Universidad de Málaga
Correo electrónico: echica@magisteriolalinea.com

Resumen

La contaminación, deforestación y agotamiento de los recursos son, entre otras, las consecuencias de la actual crisis ecológica, que tiene efectos nocivos sobre la población y al mismo tiempo está provocando una clara preocupación por el medio ambiente. Junto a ello, aparece una creciente concienciación ecológica en sectores sociales cada vez más amplios, con iniciativas voluntarias que buscan la protección y mejora del medio y producen un impacto social y ambiental de gran importancia. Los voluntarios son el recurso fundamental con el que cuentan las diferentes organizaciones, y en éstas, la mejora de su gestión interna y la implicación de los voluntarios se convierten en retos fundamentales, dado que esta implicación se concreta en un compromiso y permanencia de los voluntarios, que al mismo tiempo está condicionada por su satisfacción. En nuestra investigación, partimos de este contexto y nos planteamos validar a través del Análisis Factorial Confirmatorio una herramienta de fiabilidad satisfactoria y con una estructura factorial parsimoniosa, construida para evaluar la calidad en organizaciones de voluntariado ambiental. Consideramos que reúne los requisitos metodológicos necesarios para estimar la satisfacción de los voluntarios, posibilitando una evaluación de la calidad.

El deterioro del medio ambiente debido a los diferentes problemas locales y globales, como la contaminación y la destrucción de los recursos naturales, que comprometen la salud de los ecosistemas y del planeta en su conjunto, está claramente provocada por la relación que los seres humanos han establecido con el medio a lo largo de su historia. Las consecuencias de esta crisis ecológica se manifiestan en forma de contaminación, deforestación, agotamiento de los recursos y sus efectos nocivos sobre las personas, entre otros, son fenómenos que provocan una clara preocupación por el medio ambiente y una creciente concienciación ecológica en sectores sociales cada vez más amplios. El medio ambiente se convierte en meta y valor, que consiste en: alcanzar una calidad de vida ambiental.

La preocupación de los gobiernos ante este tema, se expresa en las reuniones que mantienen para buscar soluciones, como por ejemplo la Cumbre de la Tierra en Río, o la Cumbre sobre cambio climático en Kyoto entre otras, en las que firman acuerdos que no terminan de cumplir.

En este escenario aparece el concepto de *desarrollo sostenible*, cuya definición más aceptada se recoge en el Informe Brundtland (1987) como *la única vía posible para continuar creciendo sin agotar los recursos del planeta.*

También surgen diversas disciplinas implicadas en la búsqueda de soluciones, como son la Psicología ambiental y la Educación Ambiental, junto a grupos y personas que intervienen de forma activa y crítica para conservar, proteger y mejorar el medio ambiente, constituyéndose el voluntariado ambiental.

Estas iniciativas voluntarias, traducidas en proyectos sociales y ambientales, pueden producir un impacto social y ambiental de gran importancia. Así pues, la labor voluntaria tiene un papel significativo en el cuidado del medio ambiente y en la búsqueda de acciones de desarrollo local particularmente. Los voluntarios son el recurso fundamental con el que cuentan las diferentes organizaciones, y en éstas, la mejora de su gestión interna y la implicación de los voluntarios se convierten en retos fundamentales.

Al hablar de la implicación de los voluntarios/as no podemos pasar por alto que ésta se concreta en el compromiso y la permanencia de los mismos en las asociaciones, entidades e instituciones donde participan. Así mismo, dicha permanencia, está condicionada al mismo tiempo por su satisfacción.

Teniendo en cuenta la importancia social del voluntariado, y también la dificultad para evaluar las acciones que se realizan, es importante incorporar procesos de mejora en la gestión de las entidades voluntarias y estrategias que favorezcan la satisfacción de los voluntarios con la actividad, de forma que se favorezca la captación y fidelización de los voluntarios/as.

Respecto a este tema, Dávila y Chacón (1991) proponen el Modelo Básico para explicar mediante una estructura de relaciones que el compromiso con la organización predice la permanencia en diferentes tipos de voluntarios a través de la intención de continuar, e incluye la satisfacción como se muestra en la figura 1.

Dada la significatividad social del voluntariado, si queremos trabajar la permanencia en una organización, este *modelo básico* nos desvela la importancia de establecer estrategias que ayuden a incrementar la satisfacción con la actividad, que en última instancia, predice la permanencia. La satisfacción es una medida adicional, percibida a través de la evaluación de la experiencia y relativa a la calidad del producto o servicio.

Desde este planteamiento queremos ofrecer una herramienta o cuestionario *Inventario de Calidad en Programas de Voluntariado Ambiental* (ICPVA) útil en la gestión de programas, que evalúa la calidad de los mismos, incorporando una

metodología participativa de los voluntarios, desde un compromiso e implicación con la organización. Entendemos que la calidad se vincula con la satisfacción de los voluntarios, y ésta última influirá en su permanencia y compromiso dentro de la organización según confirman los modelos teóricos.

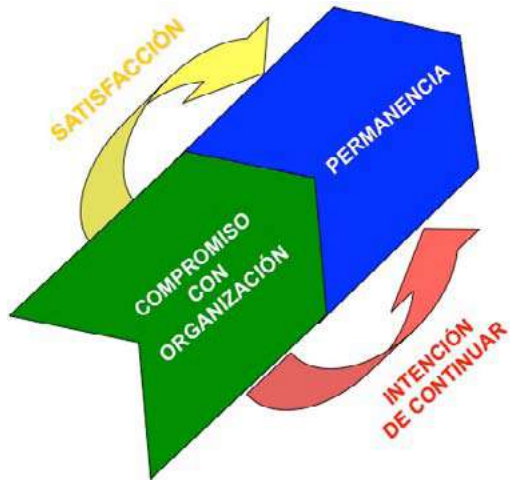


Figura 1. Esquema de relaciones según el Modelo Básico de Dávila y Chacón

En este punto es en el que se enmarca nuestro trabajo en el que nos planteamos un objetivo general consistente en elaborar una herramienta para evaluar la calidad de programas en organizaciones de voluntariado ambiental, y como objetivo específico *validar dicha herramienta con un análisis factorial confirmatorio,* que centra nuestra investigación y forma parte de otra más amplia.

MÉTODO

Respecto al material utilizado tenemos la herramienta o cuestionario Inventario de Calidad en Programas de Voluntariado Ambiental (ICPVA v.1.0) en su versión 1. Se ha utilizado el paquete estadístico SPSS v.15.0, el programa Lisrel 8.30 y Prelis 2.30.

Los participantes han sido 102 universitarios con edades comprendidas entre los 18 y 23 años que colaboran de forma habitual y/o puntual en programas de voluntariado.

El cuestionario antes dicho consta de un total de 46 ítems distribuidos en cinco escalas referidas a las tareas y relaciones de grupo (13 ítems), materiales (8 ítems), personal responsable (6 ítems), transporte (6 ítems) y comida (6 ítems).

Las respuestas se emiten en un continuo del uno al cinco e incorpora la recogida de datos relativos a la edad y género.

PROCEDIMIENTO

Respecto al procedimiento, además de elaborar el cuestionario, se realizó un *análisis factorial exploratorio* (AFE) y un *análisis de fiabilidad*. Previo al AFE, se realizaron las pruebas encaminadas a determinar la pertinencia de dicho análisis y una de ellas es el examen de la matriz de correlación. De los resultados de estas pruebas, vimos que los criterios de pertinencia se cumplían.

El siguiente paso consistió en validar nuestra herramienta con un análisis factorial confirmatorio (AFC) utilizando como procedimiento de cálculo para el ajuste del modelo el de máxima verosimilitud. Sabemos que este tipo de análisis corrige deficiencias del AFE analizando las relaciones entre un conjunto de variable observadas y una o más variable latentes; es una estrategia útil en el ámbito de la prueba de hipótesis y confirmación de teorías. En este análisis, el índice de ajuste por excelencia es÷2.

RESULTADOS

Del examen de la matriz de correlaciones, se vio que las variables estaban intercorrelacionadas, siendo éste un requisito para el análisis factorial exploratorio sea pertinente. En la tabla 1 presentamos los resultados del AFE, técnica estadística multivariante, cuyo objetivo es sacar a la luz la estructura subyacente en la matriz de datos y ver interrelaciones entre variables. Con este análisis se pretendía calcular las dimensiones latentes o factores. El porcentaje de la varianza explica el conjunto de ítems representativos y el porcentaje de varianza que puede atribuirse al error. Así mismo, Alpha mide la consistencia interna de los factores obtenidos y del total de la escala.

Tabla 1. Resultados Análisis Factorial

	Nº Factores	% varianza	Fiabilidad
Escala 1	4	63,03%	0,80
Escala 2	2	61,08%	0,74
Escala 3	4	68,41%	0,88
Escala 4	1	55,56%	0,83
Escala 5	1	67,40%	0,90

De estos resultados podemos decir que se obtuvo una herramienta de fiabilidad satisfactoria y con una estructura factorial parsimoniosa.

El siguiente paso consistió en validar nuestra herramienta con un análisis factorial confirmatorio (AFC) utilizando como procedimiento de cálculo para el ajuste del modelo el de máxima verosimilitud. Sabemos que este tipo de análisis corrige deficiencias del AFE analizando las relaciones entre un conjunto de variable observadas y una o más variable latentes; es una estrategia útil en el ámbito de la prueba de hipótesis y confirmación de teorías. En este análisis, el índice de ajuste por excelencia es÷2.

En nuestro estudio obtuvimos los índices de ajuste y error, además de los valores de Chi-cuadrado y grados de libertad. De los resultados se extrajo que el Indice de bondad de ajuste (GFI) tiene buen ajuste dado que sus valores son próximos a 1. También devuelve un buen ajuste el índice ajustado de bondad (AGFI) con valores superiores a 0.9. El índice de ajuste comparado (CFI) es adecuado obteniendo valores superiores a 0,95, al igual que el índice de ajuste no normado (NNFI) con valores entre 0,92 y 1 en todas las escalas

Los valores del resíduo cuadrático medio (RMR y RMSR) están todos por debajo de 0,1 a excepción de la escala 2 (valor 0.11), por lo que podemos considerar un ajuste aceptable para las escalas 1, 3, 4 y 5 con valores próximos a cero. El error de aproximación cuadrático medio (RMSEA) tiene valores 0,0 e indica un buen ajuste.

Respecto a los grados de libertad, que evalúan el ajuste global o en qué medida el modelo se ajusta a la población, oscilan entre 9 (escala 4 y 5), 19 (escala 2) y 59 (escala 1 y 3).

En nuestro estudio, los valores Chi-cuadrado son altos, a excepción de la escala 4. Pero como afirma Arias, este índice raramente es utilizado como prueba única o concluyente de bondad de ajuste de un modelo, ante lo que deben tenerse en cuenta otros índices.

La evaluación del ajuste final se acompañó de un análisis de la fiabilidad y la validez. Y a modo de ejemplo detallamos los resultados correspondientes a la escala 1. En ella, la fiabilidad compuesta obtiene valores superiores a 0.67. Por tanto, los indicadores de satisfacción, implementación, utilidad y oferta (factores de dicha escala) son una medida fiable del constructo.

También en esta escala 1, la varianza media extractada, medida complementaria a la fiabilidad compuesta, obtiene valores superiores a 0.50 para los factores de satisfacción y oferta, lo que ofrece confianza en las variables, aunque los valores obtenidos para los otros dos factores (implementación y utilidad) son algo inferiores, planteándonos la revisión de estos ítems.

En cuanto a la validez convergente, indicador del grado en que dos o más intentos de medir el mismo concepto están de acuerdo entre sí, los valores de t en esta escala son superiores a 1.96, lo que nos proporciona evidencia de la validez de los indicadores utilizados para medir el constructo.

En cuanto a la validez discriminante de esta escala, el valor de la varianza media extractada en cada variable es mayor al cuadrado de su correlación. También los residuos estandarizados de dicha escala presentan un ajuste bueno. Al mismo tiempo, el path de análisis confirma que el modelo consta de los cuatro factores extraídos en el AFE (satisfación-mejora, implementación, utilidad y oferta), y que están intercorrelacionados. Las covarianzas entre los mismos oscilan entre 0,0 y 0,62. En síntesis, respecto a la adecuación de los parámetros-indicadores estimados por este modelo consideramos que son razonables, dado que no aparecen varianzas negativas.

CONCLUSIONES

A la vista de los resultados, y sabiendo que estos no podemos contrastarlos con otros trabajos, porque las herramientas que existen no tienen especificadas las propiedades psicométricas, podemos decir que hemos obtenido una herramienta con fiabilidad y validez satisfactoria, su estructura factorial es parsimoniosa y reúne los requisitos metodológicos necesarios para estimar la satisfacción de los voluntarios, aportando información útil en aspectos relacionados con programas de voluntariado ambiental, así como posibilitando una evaluación de la calidad.

Aún cuando necesitamos mejorar la herramienta en los aspectos antes descritos, podemos decir que con pequeños desajustes en algunas escalas, ésta es una herramienta óptima.

REFERENCIAS

Batista Foguet, J.M. y Coenders Gallart, G. (2000). *Modelos de ecuaciones estructurales*. Salamanca: Ed. Hespérides.

Chacón Moscoso, S., Anguera, M.T. y López Ruiz, J. (2000). Diseños de evaluación de programas: bases metodológicas. *Psicothema, 12* (Supl.n.º 2), 127-131.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: *theory of generalizability for scores and profiles. New York: John Wiley and Sons.*

Dávila, C y Chacón, F. (2004). Factores psicosociales y tipo de voluntariado. P*sicothema. 16,* (4), 639-645.

Dávila, C y Chacón, F. (2007). Predicción de la permanencia del servicio voluntario: una oferta alternativa básica. *The Spanish Hournal of Psychology. 10*, (1), 115-121.

Fernández-Ballesteros, R. (1995b). Cuestiones conceptuales básicas en evaluación de programas. En R. Fernández-Ballesteros (Ed.): *Evaluación de programas. Una guía práctica en ámbitos sociales, educativos y de salud* (pp. 21-47). Madrid: Síntesis

Hernández Mendo, A. (2001). Un cuestionario para evaluar la calidad en programas de actividad física. Revista *de Psicología del Deporte, 10, 179-196.*

Hernández Mendo, A. (2001).Un cuestionario para evaluar la calidad en programas de actividad física. *Revista de Psicología del Deporte*, *10*(2), 179-196

Jöreskog, K.G. y Sijrbom, D. (1993). LISREL 8: *Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.

RECURSOS EFICIENTES PARA MEDIDAS REPETIDAS NATURALMENTE NECESARIAS

Coordinador: Paula Fernández García
Universidad de Oviedo

No cabe ninguna duda de que los diseños de medidas repetidas despiertan un enorme interés. Para muestra un botón. Ni que decir tiene que en lo que a la Psicología se refiere, todos aquellos que hemos participado con cierta regularidad en los anteriores XI congresos de Metodología de las Ciencias del Comportamiento y de la Salud, entre otros, venimos encontrando cabida en simposios, mesas de comunicaciones y salas de Póster para el estudio del método o la aplicación del método que es connatural y que socorre a las medidas repetidas. Estos diseños de naturaleza longitudinal son trasversales sin embargo, a saber, trasversales porque están omnipresentes en todo tipo de investigación que busca obtener inferencias causales, ya sea la investigación experimental, cuasi-experimental o no experimental; trasversales porque son los más utilizados en todos los campos de investigación, epidemiología, medicina, farmacia, biología, agricultura, marketing, ecología…; trasversales porque metodológicamente son abordados analíticamente desde múltiples perspectivas: univariada, multivariada, mixta, remuestreo, ecuaciones estructurales, gráficamente…, y así podríamos seguir describiendo esta condición que también viene a definirlos. Aunque es verdad que en función de cuál sea la unidad de análisis sometida a estudio y qué se pretenda estudiar (por ejemplo, si son personas, a veces el interés está centrado en el estudio de la evolución y cambio a través del tiempo de alguna variable ya sea comportamental, cualquiera en su amplio espectro, u orgánica; otras el interés está centrado en la puesta a prueba de la eficacia de varios tratamientos, y que necesariamente se llevan a cabo con las mismas personas por carecer de muestra suficiente; otras veces es necesario estudiar la existencia de efectos residuales cuando necesariamente se tienen que aplicar varios tratamientos a las mismas personas, e incluso estos efectos puede que sean deseables..., etc), el ámbito donde se lleva a cabo el estudio (laboratorio, condiciones naturales) y la dirección del estudio (si es prospectiva o histórica), el analista de los datos puede esperar encontrarse con dificultades concretas. Sin embargo hay algunos problemas que son democráticos y por lo tanto susceptibles de aparecer siempre. Son los siguientes: autocorrelación de las puntuaciones, la pérdida de algunos datos para alguna de las unidades experimentales o la mortandad experimental definitiva en algún momento del tiempo, y disponer de grupos de unidades experimentales no equilibrados. Estos problemas afectan a la detección de los efectos de las variables de interés en el estudio y sus interacciones. Y aunque es verdad que

el riesgo se puede reducir apelando a una cuidadosa planificación de la investigación y recogida de datos, no es con frecuencia suficiente, y el recurso que nos queda es aplicar correctamente la técnica de análisis de datos más oportuna y eficiente. El interés de este simposio es exponer el estudio sobre recursos existentes para que las investigaciones de medidas repetidas sean más eficientes.

PALABRAS CLAVE: Medidas repetidas, Potencia, Robustez, Desequilibrio en los grupos, Procedimientos paramétricos y no paramétricos.

ANÁLISIS DE DATOS LONGITUDINALES INCOMPLETOS USANDO MODELOS MARGINALES

G. Vallejo[1], M. P. Fernández[1], P. E. Livacic-Rojas[2], E. Tuero-Herrero[1], J. F. García[3] y M. Castillo-Fuentes[3]

[1] Universidad de Oviedo

[2] Universidad de Santiago de Chile

[3] Universidad de Valencia

Correo electrónico: gvallejo@uniovi.es.

Resumen

Dado que los datos faltantes costituyen un problema generalizado en muchas aplicaciones psicológicas en el mundo real, hemos estudiado el impacto que la pérdida de sujetos tiene en la sensibilidad de varios enfoques de promedio poblacional. En concreto, se han examinado el modelo de regresión con patrones de covarianza verdadero (MPC-T) y no estructurado (MPC-U), el método de las ecuaciones de estimación generalizada combinado con la técnica de imputación múltiple (MI-GEE) y el método GEE ponderado (GEE). Recientemente, Vallejo, Fernández, Livacic-Rojas y Tuero-Herrero (2011) han encontrado que bajo el denominado mecanismo de abandono MAR, el enfoque MI-GEE era el método más robusto de los reseñados, aunque empleando moderados tamaños de muestra, los métodos MPC-T y MPC-U también controlaban adecuadamente las tasas de error. Por lo que respecta a la potencia, los resultados encontrados en la presente comunicación ponen de relieve que los métodos basados en la verosimilitud eran más potentes que los basados en la cuasi-verosimilitud y su superioridad fue a menudo considerable. Por otra parte, observamos que poca o ninguna potencia era sacrificada cuando se empleaba el método MPC-U en lugar de la MPC-T, aunque ambos métodos tenían menos potencia cuando los sujetos abandonaban prematuramente el estudio.

En las investigaciones psicológicas resulta usual que los sujetos se subdividan en grupos de acuerdo a los niveles de un factor de tratamiento, o bien de clasificación, y se midan en más de una ocasión temporal. Debido a que las reiteradas observaciones se registran desde la misma unidad, una característica clave de estos estudios es la posible dependencia de los errores. De ahí que los métodos que permiten modelar la correlación entre las respuestas sean cada vez más empleados. Entre estos, destacan el modelo de regresión con patrón de covarianza estructurado (CPM; Jennrich y Schluchter, 1986) y el modelo lineal de efectos mixtos (LMM, Laird y Ware, 1982). Estos enfoques, más que asumir estructuras de covarianza

excesivamente parcas (p.e., la matriz de simetría compuesta típica del enfoque ANOVA de medidas repetidas) o completamente generales (p.e., la matriz no estructurada típica del enfoque MANOVA), tratan de buscar un equilibrio entre los criterios de flexibilidad y parsimonia.

Recientemente, Vallejo, Arnau, Bono, Fernández y Tuero (2010) y Vallejo, Fernández, Livacic-Rojas y Tuero (2011a) han constatado que las herramientas de selección de modelos resultan de utilidad limitada cuando estructuras de covarianza complejas se combinaban con muestras pequeñas y cuando los datos se obtenían a partir de distribuciones sesgadas. En estos casos, más que tratar de ajustar la mejor estructura de covarianzas, puede ser más conveniente asumir algún patrón de covarianzas de antemano (Kowalchuk, Keselman, Algina, y Wolfinger, 2004; Mallinckrodt et al., 2004).

Otro rasgo característico de los estudios de repetidas medidas es el problema de de los datos faltantes. En muchas aplicaciones psicológicas en el mundo real, la pérdida de datos resulta inevitable, generalmente por abandono de los sujetos incluidos en el estudio. Basándose en la distribución del mecanismo de pérdida, Rubin (1976) define una jerarquía de tres procesos diferentes de no respuesta, a saber: Datos perdidos completamente al azar (MCAR), datos perdidos al azar (MAR) y datos perdidos no ignorables (MNAR). Aunque en la realidad es casi imposible especificar correctamente el mecanismo de la ausencia de datos, varios estudios, incluyendo los de Little y Rubin (2002) y Mallinckrodt et al. (2004), han concluido que el MAR es un mecanismo realista para la mayoría de las aplicaciones prácticas, debido a que los abandonos están frecuentemente relacionados con las respuestas dadas anteriormente en el estudio.

Existen varias alternativas para tratar de sortear las limitaciones del modelo lineal clásico, incluyendo los métodos paramétricos basados en la verosimilitud, los métodos basados en imputar múltiples (MI) valores para cada dato perdido y los métodos semi-paramétricos basados en la regresión. Los métodos basados en la verosimilitud (p.e., los tests CPM y LMM) constituyen la solución más extendida para analizar datos normales incompletos, debido a su validez bajo MAR. Los métodos MI también son válidos bajo MAR e implican generar varias versiones plausibles de los valores perdidos, analizar cada conjunto de datos imputados separadamente usando técnicas estándar y combinar los resultados de acuerdo a las reglas de Rubin (1987). Mientras que los métodos semi-paramétricos, tales como las ecuaciones de estimación generalizada (GEE; Liang & Zeger, 1986) y las ecuaciones de estimación generalizada ponderadas (WGEE; Robins, Rotnitzky & Zhao, 1995), constituyen una alternativa atractiva a los métodos paramétricos cuando los datos son MAR y se viola el supuesto de normalidad.

Aunque se han realizado numerosos estudios para evaluar el desempeño de los enfoques reseñados (Beunckens, Sotto y Molenberghs, 2008; DeSouza, Legedza y Sankoh, 2009; Mogg y Mehrotra, 2007; Padilla y Algina, 2007; Vallejo, Fernández, Livacic-Rojas y Tuero, 2011b), la mayor parte de los mismos se han centrado en comparar el desempeño en términos de especificidad, sesgo y precisión. Por con-

siguiente, el objetivo de la presente comunicación radica en investigar el rendimiento de los métodos paramétricos (CPM basado en la verdadera estructura de covarianza y CPM basado en una estructura de covarianza asumida no estructurada, en adelante CPM-T y CPM-U, respectivamente) y semi-paramétricos (MI-GEE y WGEE) con respecto a su sensibilidad para detectar cambios en el tiempo.

MÉTODO

Para evaluar la sensibilidad de los procedimientos descritos llevamos a cabo un estudio de simulación utilizando un diseño completamente al azar, en el cual sujetos fueron distribuidos aleatoriamente en dos grupos de tratamiento y medidos en ocho ocasiones a lo largo del tiempo. Las variables manipuladas en el estudio fueron las siguientes:

1. *Patrones de covarianza.* Los patrones utilizados para generar los datos fueron cuatro, a saber: Autoregresivo homogéneo de primer orden (AR), autoregresivo heterogéneo de primer orden (ARH), Toepliz heterogéneo (TOEPH) y no estruturado (UN). Los detalles de estos patrones se muestran en la Tabla 1.
2. *Tamaños totales de muestra*. Tres tamaños de la muestra fueron considerados, $n = 30$, $n = 60$ y $n = 120$. En ausencia de pérdida de sujetos, el tamaño muestral más pequeño proporcionaba una potencia prospectiva del 50% para la interacción tratamiento × tiempo, el tamaño de la muestra intermedio del 70% y el tamaño de la muestra mayor del 90%. El tamaño de los grupos fue: (a) 10, 20 ($n = 30$), (b) 20, 40 ($n = 60$), y (c) 40, 80 ($n = 120$).
3. *Forma de la distribución de la variable de medida*. Para investigar el efecto que ejerce la forma de la distribución en el desempeño de las técnicas analíticas, generamos datos desde distribuciones normales y no normales mediante las distribuciones *g* y *h* introducidas por Tukey. Además de la distribución normal también investigamos otras dos: (a) una distribución que tiene el mismo grado de sesgo y exceso de curtosis que la doble exponencial o Laplace; y (b) una distribución que tiene el mismo grado de sesgo y de exceso de curtosis que la exponencial.

El desempeño de los métodos anteriormente citados se investigó cuando los abandonos seguían un patrón monótono MAR y la tasa abandonos acumulados fue del 50%. Se realizaron 5000 réplicas de cada condición estudiada con un nivel de significación del 5% mediante un MACRO escrito en lenguaje SAS/IML.

Tabla 1. Tasas de potencia (%) para datos completos e incompletos homogéneos

	AR(1)			ARH(1)			TOEPH			UN		
Método	$n = 30$	$n = 60$	$n = 120$	$n = 30$	$n = 60$	$n = 120$	$n = 30$	$n = 60$	$n = 120$	$n = 30$	$n = 60$	$n = 120$
Datos normales con observaciones completas												
CRM-T	50.2	70.8	89.4	50.9	70.8	89.8	50.3	70.5	90.1	50.9	70.0	89.1
CRM-U	44.7	68.2	88.0	46.8	66.4	89.1	48.4	69.1	89.7	50.9	70.0	89.1
Datos Laplace con observaciones completas												
CRM-T	54.2	74.0	89.9	54.6	70.9	88.3	56.8	74.5	89.1	53.8	73.6	91.4
CRM-U	49.6	70.9	88.6	49.4	67.4	87.6	53.8	72.8	88.8	53.8	73.6	91.4
Datos sesgados con observaciones completas												
CRM-T	52.1	72.4	88.8	55.3	72.1	89.3	54.4	72.8	89.6	58.2	73.5	89.3
CRM-U	51.4	71.9	88.9	52.2	71.7	89.4	55.2	72.7	90.1	58.2	73.5	89.3
Datos normales con pérdidas MAR												
CRM-T	38.6	58.3	80.2	38.9	59.4	81.0	37.9	58.1	80.1	33.4	53.4	73.8
CRM-U	34.0	54.2	78.0	33.7	57.7	79.9	36.6	57.9	79.2	33.4	53.4	73.8
WGEE	28.9	36.0	48.0	30.3	35.9	47.5	34.0	37.9	44.7	22.9	29.6	35.9
MI-GEE	34.0	53.3	78.1	30.0	54.2	76.6	31.1	50.0	73.2	23.8	39.2	55.6
Datos Laplace con pérdidas MAR												
CRM-T	44.5	63.5	81.2	45.6	64.0	82.9	46.0	65.9	83.7	39.6	66.7	83.4
CRM-U	43.4	60.0	81.9	46.0	61.0	81.8	43.1	63.8	82.2	39.6	66.7	83.4
WGEE	33.4	42.3	44.6	37.5	43.5	49.0	40.4	45.4	**49.6***	27.4	38.4	39.3
MI-GEE	37.0	59.6	79.5	37.4	55.1	78.3	38.4	56.4	75.9	28.7	50.5	64.7
Datos sesgados con pérdidas MAR												
CRM-T	32.2	48.3	70.0	33.7	52.2	77.7	35.8	49.7	68.1	32.8	47.9	62.5
CRM-U	30.7	47.2	67.9	32.8	52.3	71.9	35.8	50.2	69.3	32.8	47.9	62.5
WGEE	29.7	43.1	58.2	37.0	48.2	56.4	42.7	48.8	59.5	24.5	31.0	42.2
MI-GEE	27.2	45.8	69.3	26.1	45.4	64.3	26.9	40.9	58.1	22.7	32.2	42.0

Nota: Los valores en negrita corresponden a condiciones bajo las cuales los enfoques no controlaban las tasas de error.

RESULTADOS DEL ESTUDIO

En la Tabla 1 aparece tabulado el porcentaje de veces que los métodos examinados rechazaban correctamente la hipótesis nula referida a la interacción de los grupos con las ocasiones de medida. Los datos denotan el porcentaje promedio de rechazos correctos a través de la variable igualdad/desigualdad del tamaño de los grupos. Globalmente, los resultados ponen de relieve lo siguiente:

- Para datos normales con observaciones completas, los resultados en la Tabla 1 indican que el método CPM-T fue ligeramente más potente que el método CPM-U cuando $n = 30$. Sin embargo, las diferencias de potencia desaparecían cuando $n > = 60$. Bajo esta situación, las tasas de potencia promedio correspondientes a los métodos CPM-T y CPM-U fueron 70.2% y 68.4%, respectivamente.

- Para datos no-normales con observaciones completas, las diferencias entre los enfoques CPM-T y CPM-U también fueron pequeñas. En concreto, las tasas de potencia promedio obtenidas con el enfoque CPM-T fueron 72.3% (distribución tipo exponencial) y 72.6% (distribución tipo Laplace), mientras que los porcentajes obtenidos con el enfoque CPM-U fueron 72.1% (distribución tipo exponencial) y 70.6% (distribución tipo Laplace).
- Para datos normales bajo el mecanismo de pérdida MAR, los resultados de la Tabla 1 indican que las potencias de los métodos CPM-T y CPM-U eran casi idénticas, aunque contrariamente a lo que sucedía cuando no había desgaste de muestra, en este caso las tasas de potencia empíricas eran sustancialmente más pequeños que las teóricas (los valores medios fueron 57.8% y 56.0%, respectivamente). A su vez, el método MI-GEE era más potente que el método WGEE. Específicamente, el método MI-GEE rechazaba la hipótesis nula en el 50% de los casos, mientras que el método WGEE lo hacía en el 36%.
- Para datos no-normales bajo el mecanismo de MAR, las diferencias de potencia entre los enfoques CPM-T y CPM-U también fueron pequeñas y muy afectadas por la pérdida de datos. En particular, las tasas de potencia promedio fueron 50.9% - 50.1% (distribución tipo exponencial) y 63.9% - 62.7% (distribución tipo Laplace) para los métodos CPM-T y CPM-U, respectivamente. A su vez, cuando la distribución subyacente a los datos era del tipo Laplace, el método de MI-GEE era más potente que el método WGEE (los valores promedio fueron 55.1% y 40.8%, respectivamente), mientras que cuando la distribución subyacente a los datos era del tipo exponencial, ambos métodos producían tasas similares (los valores medios fueron 41.7% y 43.4%).

CONCLUSIONES Y RECOMENDACIONES

Los resultados revelaron que las tasas de potencia correspondientes a los métodos basados en la verosimilitud eran más grandes que las tasas de la potencia de los métodos basados en las ecuación de estimación generalizada, y en algunas condiciones sustancialmente más grandes. Este patrón de resultados se observó tanto bajo el mecanismo de pérdida MAR, tanto monótona como intermitente, independientemente de la forma de distribución de la población. Aunque en un estudio previo (Vallejo, Fernández, Livacic-Rojas & Tuero, 2011) hemos encontrado que el método MI-GEE controla mejor que sus competidores las tasa de errores, por desgracia, resultó ser menos potente. También se encontramos que el método WGEE, a pesar de que el método inflar las tasas de error, tenía problemas de potencia para detectar efectos de interacción. Por lo tanto, a la luz de nuestros resultados y los reportados por Mogg y Mehrotra (2007) y Souza et al. (2009), se recomienda utilizar el método WGEE con la debida precaución.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado mediante el proyecto de investigación concedido por el MCI (Ref.: PSI-2008-03624).

REFERENCIAS

Beunckens, C., Sotto, C., & Molenberghs, G. (2008). A Simulation study comparing weighted estimating equations with multiple imputation based estimating equation for longitudinal binary data. *Computational Statistics & Data Analysis, 52*, 1533-1548.

DeSouza, C., Legedza, A. T., & Sankoh, A. J. (2009). An overview of practical approaches for handling missing data in clinical trials. *Journal of Biopharmaceutical Statistics, 19*, 1055-1073.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated–measures models with structured covariance matrices. *Biometrics, 42*, 805-820.

Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational & Psychological Measurement, 64*, 224-242.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963-974.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13-22.

Little, R. J. A., & Rubin, D. B. *(*2002*). Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley.

Mallinckrodt, C. H., Kaiser, C. J., Watkin, J. G., Detke, M. J., Molenberghs, G., & Carroll, R. J. (2004). Type I error rates from likelihood based repeated measures analyses of incomplete longitudinal data. *Pharmaceutical Statistics, 3*, 171–186.

Mogg*, R., &* Mehrotra, D.(2007). Analysis of antiretroviral immunotherapy trials with potentially non-normal and incomplete longitudinal data. *Statistics in Medicine, 26*, 484-497.

Padilla, M. A., & Algina, J. (2007). Type I error rates of the Kenward-Roger adjusted degree of freedom *F*-test for a split-plot design with missing values. *Journal of Modern Applied Statistical Methods*, *6*, 66-80.

Robins, J., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. *Journal of the American Statistical Association*, *90*, 106-121.

Rubin, D. B. *(*1987*). Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Vallejo, G., Arnau, J., Bono, R., Fernández, P., & Tuero, E. (2010). Selección de modelos anidados para datos longitudinales usando criterios de información y la estrategia de ajuste condicional. *Psicothema, 22*, 323-333.

Vallejo, G., Fernández, P., Livacic-Rojas, P., & Tuero, E. (2011a). Selecting the best covariance pattern regression model with missing data. *Behavior Research Methods, 43*, 18-36.

Vallejo, G., Fernández, P., Livacic-Rojas, P., & Tuero, E. (2011b). Comparison of modern methods for analyzing unbalanced repeated measures data with missing values. *Multivariate Behavioral Research*, *46* (*6*), 1-38.

ESTUDIO COMPARATIVO DEL COMPORTAMIENTO DE DOS SELECTORES DE ESTRUCTURAS DE COVARIANZA Y SUS TASAS DE ERROR PARA ANALIZAR DATOS CON DISEÑOS SPLIT PLOT

Pablo Livacic-Rojas[1], Guillermo Vallejo[2], Paula Fernández[2] y Ellián Tuero[2]
[1] Universidad de Santiago de Chile
[2] Universidad de Oviedo
Correo electrónico: pablo.livacic@usach.cl

Resumen

El presente trabajo analiza la frecuencia de selección de estructuras de covarianza y los niveles de error entre el criterio de Akaike (AIC) y el Modelo Correctamente Identificado. (MCI). Mediante un estudio de Simulación Montecarlo con el software SAS 9.1 se utilizó un diseño Split-Plot manipulando las variables tamaño muestral, relación entre el tamaño de los grupos y el de las matrices de dispersión, tipo de matrices de dispersión y forma de la distribución. Los resultados muestran que AIC selecciona estructuras de covarianza heterogéneas y exhibe tasas de error de tipo I más altas que MCI. Se estima necesario estudiar a futuro los niveles de potencia de los selectores en presencia de vectores con datos completos e incompletos.

Dentro de las dificultades más frecuentes para analizar datos en diseños de medidas repetidas con el modelo clásico están las puntuaciones correlacionadas en la misma unidad experimental a través del tiempo, la diferencia entre las unidades respecto a la cantidad y el momento de las observaciones, presencia de datos heterogéneos, observaciones faltantes, covariables dependientes del tiempo y la violación de los supuestos paramétricos (Vallejo et al, 2010; Fernández, et al, 2010. ; Vallejo et al, 2011; Livacic-Rojas et al, 2010; Vallejo et al, 2008; Keselman et al 2001). A su vez, Kowalchuk et al. (2004) señalan como una dificultad inherente a los procedimientos de ANOVA y MANOVA la suposición de una estructura de covarianza subyacente, la cual, al no estimar una serie de parámetros del modelo los hace ineficientes.

El uso de modelos lineales mixtos (MLM) se ha incrementado al ser flexible para la modelización de errores correlacionados con las estructuras de covarianza parsimoniosas, lo cual, influye en la frecuencia de selección y en los niveles de potencia. Fernández et al. (2010) señalan que para hacer inferencias su uso tiene como ventajas: 1) analizar datos de vectores completos e incompletos, 2) modelar la variabilidad de los sujetos a nivel entre e intra, 3) manejar covariables

dependientes del tiempo y, 4) generalizarse a contextos multivariados. Como desventajas señalan: 1) problemas de AIC y el criterio de información Bayesiano de Schwarz (BIC) para identificar el verdadero proceso subyacente de los datos cuando la estructura de covarianza es compleja, y 2) bajo rendimiento para muestras pequeñas.

Vallejo et al. (2010) señalan que el rendimiento de los criterios de información ha sido empíricamente evaluado a través de la MLM en tres contextos analíticos diferentes:

1. Capacidad para seleccionar el modelo correcto dada una estructura de covarianza particular (Gurka 2006; Wang y Schaalje, 2009).
2. Capacidad para seleccionar la estructura de covarianza cuando la media del modelo se conoce (Keselman et al 1998; Ferron et al 2002; Gómez et al 2005; Gurka, 2006; Vallejo et al, 2008).
3. Capacidad para seleccionar al mismo tiempo la estructura media y la covarianza (Gurka, 2006; Vallejo et al, 2011).

Vallejo et al. (2011) sugieren que no existe consenso general sobre el mejor criterio para seleccionar estructuras de covarianza dado que el rendimiento de cada uno depende de las condiciones de la investigación. Ferron et al. (2002) señalan que AIC selecciona correctamente la verdadera estructura de covarianza en un 79% de las ocasiones analizadas y BIC el 66%. A su vez, Keselman et al. (1998) encontraron que AIC escoge adecuadamente el 47% y BIC el 35%. Vallejo et al. (2010) indican que AIC lo realiza el 68% y BIC el 48%. En otro estudio, Vallejo et al. (2011) plantean que AIC selecciona correctamente el 44.5% y BIC 43.5%. Por último, Gómez et al. (2005) encontraron que los dos criterios las seleccionan el 22%.

Los estudios señalados compararon la eficacia de la AIC y BIC junto a otros selectores, encontrándose que la selección de la estructura de covarianza mejora cuando aumenta el tamaño de la muestra y la matriz es más simple. En función de lo expuesto, los objetivos del presente estudio fueron: 1) comparar la frecuencia de la selección de estructuras de covarianza para los criterios AIC y MCI y 2) comparar las tasas de error tipo I para la los efectos entre e intra sujetos y de la interacción.

MÉTODO

Se evaluó el desempeño de la selección de Akaike mediante un estudio de simulación con el software estadístico SAS 9.1. La selección de entre 12 posibles estructuras de covarianza fueron simetría compuesta (SC), no estructurada (NE), de primer orden autorregresivo [AR (1)] , Huynh-Feldt (HF), simetría compuesta heterogéneos (SCH), autorregresiva heterogénea de primer orden [ARH (1)], el modelo lineal de coeficientes aleatorios con heterogeneidad entre los grupos (CAL), matriz de covarianza no estructurados con en un mismo sujeto y entre el

grupo- heterogeneidad (NEH), Huynh-Feldt estructura esférica con la heterogeneidad entre los grupos (HFH), heterogéneos autorregresivos (ARHJ), y coeficientes aleatorios heterogéneos (CAJ). El diseño de la investigación fue Split Plot con un factor de entre sujetos (p = 3) y uno intra-sujetos factor (k = 4) y, se manipularon cuatro variables: (a) tamaño de la muestra total, (b) la relación entre el tamaño del grupo y el tamaño de la matriz de dispersión, (c) el tipo de matriz de dispersión, y (d) la forma de la distribución.

Las tasas de error Tipo I fueron evaluadas con criterio liberal de Bradley (Bradley, 1978) con el intervalo $\left(2.5 \leq \hat{\alpha} \leq 7.5\right)$ y un nivel de significación del 5%, considererándose un procedimiento robusto si la tasa de error empírico se situaba en el intervalo $\left(0.025 \leq \hat{\alpha} \leq 0.075\right)$ bajo cualquier condición.

RESULTADOS

La Tabla 1 muestra que AIC selecciona la estructuras NE, CA, y ARH en un 2.06%, 0.67% y 0.69% de las ocasiones analizadas, respectivamente. En la misma línea para las matrices heterogéneas se selecciona NEH, ARHJ, CAH, y HFH en 34.71%, 28,69%, 27.45% y 5.82% respectivamente. En cuanto a la estructura de covarianza original y el tamaño del grupo, se encontró que selecciona correctamente NE, CA y ARH en 4.46%, 1.63% y 1.06% y para una relación de tipo negativa.

La Tabla 2 muestra para las tasas de error de tipo I lo siguiente:

1. Del total de condiciones estudiadas, AIC las supera el límite superior el 7.41% y MCI el 2.47%.
2. Las tasas de error para lo efectos entre e intra-grupo y de la interacción fueron superadas en un 4.94% y estaba asociado con las matrices CA y NE (n= 30) para una relación negativa. Con tamaños de 45 y 60, las tasas se superaron en el 1.23% asociado a la matriz de CA y para una relación de tipo negativa.
3. El MIC supera el criterio de Bradley en el 2.47% de las condiciones de los efectos entre-grupos y se asocia con una matriz NE (n= 30) y una relación negativa entre el tamaño grupal y la estructura de covarianza.

Tabla 1. Porcentajes promediados selección estructuras covarianza con datos distribuidos normalmente y alejados de la normalidad.

N	R	ECO	SC	NE	AR1	HF	SCH	ARH(1)	CAL	NEH	HFH	ARHJ	CAH
30	=	ARH(1)	0.25	1.28	4.50	0.20	0.10	1.39	0.60	24.8	6.20	39.8	20.9
		CA	0.00	1.60	0.10	0.25	0.43	0.25	2.55	33.7	9.63	6.43	45.2
		NE	0.33	3.98	1.88	0.20	0.43	1.85	0.20	35.8	9.43	29.4	16.6
	+	ARH (1)	0.10	1.85	7.25	0.05	0.33	3.22	0.93	29.7	7.55	31.1	15.7
		CA	0.05	1.93	0.03	0.23	0.85	0.45	6.05	36.2	11.4	6.13	36.8
		NE	0.38	7.30	2.85	0.18	0.85	2.82	0.43	36.3	12.0	24.2	2.98
	–	ARH (1)	0.05	1.73	3.40	0.13	0.00	1.60	0.58	31.4	7.95	33.3	19.9
		CA	0.00	1.50	0.08	0.10	0.53	0.18	2.83	35.7	13.1	6.63	40.2
		NE	0.33	14.9	1.40	0.10	0.63	1.25	0.15	40.9	12.6	24.4	13.2
45	=	ARH(1)	0.00	0.35	0.83	0.00	0.00	0.53	0.13	24.2	1.50	54.0	18.6
		CA	0.00	0.65	0.00	0.18	0.03	0.03	0.63	13.0	22.9	4.10	**57.2**
		NE	0.00	2.45	0.33	0.08	0.10	0.45	0.10	43.7	4.30	35.6	12.9
	+	ARH (1)	0.03	0.95	2.28	0.08	0.00	1.30	0.18	26.2	2.00	49.9	26.1
		CA	0.00	1.13	0.00	0.05	0.15	0.05	1.25	36.3	3.87	11.3	**54.6**
		NE	0.08	5.60	2.40	0.05	0.28	0.90	0.03	43.4	3.70	37.1	26.0
	–	ARH (1)	0.00	0.40	0.88	0.03	0.00	0.75	0.08	26.5	2.03	50.9	16.0
		CA	0.00	0.60	0.00	0.00	0.08	0.05	0.68	34.7	5.60	4.10	**56.7**
		NE	0.00	1.58	0.25	0.05	0.20	0.38	0.03	45.6	4.93	34.4	12.1
60	=	ARH(1)	0.00	0.10	0.10	0.00	0.00	0.18	0.05	23.2	0.53	**61.2**	15.9
		CA	0.00	0.13	0.00	0.00	0.03	0.03	0.20	35.3	2.35	2.28	**59.8**
		NE	0.00	0.65	0.03	0.00	0.00	0.10	0.00	**52.6**	2.58	34.0	10.1
	+	ARH (1)	0.00	0.43	0.43	0.00	0.03	0.40	0.00	24.6	0.78	**58.4**	14.7
		CA	0.00	0.45	0.00	0.03	0.03	0.00	0.33	36.2	1.58	1.80	**59.6**
		NE	0.00	2.78	0.08	0.05	0.05	0.23	0.00	**52.7**	2.60	33.5	8.03
	–	ARH (1)	0.00	0.08	0.13	0.03	0.00	0.15	0.03	24.4	1.00	**60.5**	13.8
		CA	0.00	0.28	0.00	0.00	0.08	0.00	0.18	36.7	2.70	5.18	**57.9**
		NE	0.00	0.88	0.00	0.00	0.03	0.08	0.00	51.0	2.40	34.9	9.73

Nota: Tamaño Muestra (N), Relación entre tamaño muestra y matriz de dispersión (R), Relación nula entre tamaño muestra y matriz de dispersión (=), Relación positiva la entre tamaño muestra y matriz de dispersión (+), Relación negativa entre tamaño muestra y matriz de dispersión (-), Estructura covarianza original (ECO), Simetría Compuesta (SC), No Estructurada (NE), Autoregresiva de Primer Orden [AR(1)], Huynh-Feldt (HF), Simetría Compuesta Heterogénea (SCH), Autoregresiva Heterogénea de Orden Uno [ARH(1)], Coeficientes Aleatorios Linealea (CAL), No Estructurada Heterogénea (NEH), Huynh-Feldt Heterogénea (HFH), Autoregresiva Heterogenea (ARHJ), Coeficientes Aleatorios Heterogenea (CAH), valores altos (negrita).

Tabla 2. Porcentajes promediados tasas error de tipo I selectores estructuras covarianza datos distribuidos normalmente y alejados de la normalidad

			AIC			MCI		
N	**R**	**ECO**	**Entre**	**Intra**	**Interacción**	**Entre**	**Intra**	**Interacción**
30	=	ARH(1)	5.58	4.80	4.18	5.75	3.93	3.00
		CA	6.73	6.85	5.70	5.78	5.65	4.45
		NE	6.20	5.25	4.83	6.25	4.53	3.38
	+	ARH (1)	4.60	4.08	4.10	4.75	3.48	2.58
		CA	4.40	6.58	4.70	3.93	5.93	4.68
		NE	4.03	5.35	4.40	3.83	4.53	3.60
	–	ARH (1)	7.43	5.70	5.38	7.33	3.75	2.70
		CA	**8.35**	**8.13**	**7.85**	**7.93**	5.48	5.33
		NE	**7.78**	5.53	6.23	**7.75**	4.00	4.80
45	=	ARH(1)	6.20	4.30	4.35	6.25	3.98	3.58
		CA	7.03	6.65	5.18	6.80	6.10	4.60
		NE	5.70	5.28	4.25	5.63	5.50	4.08
	+	ARH (1)	4.45	3.33	3.40	4.73	3.98	2.93
		CA	5.00	6.58	4.20	4.60	5.45	4.08
		NE	4.70	4.60	3.75	4.75	4.60	3.85
	–	ARH (1)	7.10	5.08	4.93	7.00	3.98	3.20
		CA	**7.55**	6.45	6.80	4.85	5.55	5.20
		NE	7.03	5.80	5.00	6.85	4.98	4.43
60	=	ARH(1)	5.88	4.35	4.03	6.10	4.30	3.30
		CA	6.18	6.03	5.65	6.03	5.15	5.18
		NE	5.38	5.08	4.80	5.18	4.85	4.60
	+	ARH (1)	5.43	4.50	3.78	5.45	3.80	3.20
		CA	5.38	5.80	4.68	5.38	5.38	4.25
		NE	5.00	4.90	4.78	5.13	4.95	4.95
	–	ARH (1)	6.95	5.28	4.33	6.90	4.55	3.35
		RC	**7.90**	7.30	6.85	7.38	6.05	5.65
		UN	6.60	5.20	4.60	6.40	4.90	4.33

Nota: Tamaño Muestra (N), Relación entre tamaño muestra y matriz de dispersión (R), Relación nula entre tamaño muestra y matriz de dispersión (=), Relación positiva la entre tamaño muestra y matriz de dispersión (+), Relación negativa entre tamaño muestra y matriz de dispersión (-), Estructura covarianza original (ECO), No Estructurada (NE), Autoregresiva Heterogénea de Orden Uno [ARH(1)], Coeficientes Aleatorios (CA), Criterio Información Akaike (AIC), Modelo Correctamente Identificado (MCI), Tasa error Tipo I que exceden Criterio Badley (Negrita).

DISCUSIÓN

El presente estudio se analiza la frecuencia de selección de estructuras de covarianza y los niveles de error mediante el criterio de Akaike y el modelo correctamente identificado. Los resultados muestran que AIC selecciona en un porcentaje mayor estructuras de covarianza heterogéneas en vez de las estructuras originales, lo cual, trae como consecuencia un funcionamiento algo defectuoso al

utilizarlo dado que esto implicaría que los investigadores estimen una mayor cantidad de parámetros al subyacer matrices de dispersión más complejas de los datos. En tal sentido, los hallazgos encontrados coinciden en una muy baja medida con los realizados por Ferron et al. (2002), Keselman et al. (1998), Vallejo et al. (2010) y Vallejo et al. (2011), y lo hacen en mayor medida con el de Gómez et al. (2005). A la luz de estos resultados, se estima oportuno analizar el comportamiento de estos criterios con sus versiones corregidas tanto cuando los vectores de datos están completos e incompletos, así como también, los niveles de potencia.

NOTA DE LOS AUTORES

El presente trabajo ha sido financiado con el proyecto de investigación concedido por el Ministerio de Ciencia e Innovación de España (Ref.: PSI-2008-03624).

REFERENCIAS

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*,144–152.

Fernández, P., Vallejo, G., & Livacic-Rojas, P. (2010). Robsutez de cinco estadísticos univariados para analizar diseños Split-Plot en condiciones adversas. *Revista Latinoamericana de Psicología, 42,* 289-309.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first level error structure in two-level models of change. *Multivariate Behavior Research, 37*, 379-403.

Gómez, V.E., Schaalje, G.B., & Fellingham, G.W. (2005). Performance of Kenward-Roger method when covariance structure is selected using AIC and BIC. *Communications in Statistics-Simulation and computation, 34,* 377-392.

Gurka, M.J. (2006). Selecting the best linear mixed models under REML. *The American Statistiacian, 60*, 19-26.

Keselman, H.J., Algina, J., Kowalchuk, R.K., & Wolfinger, R.D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics- Simulation and Computation, 27*, 591-604.

Keselman, H.J. Algina, J., & Kowalchuck, R.K. (2001). The analysis of the repeated measures design: A review. *British Journal of Mathematical and Statistical Psychology, 54,* 1-20.

Kowalchuck, R.K., Keselman, H.J., Algina, J., & Wolfinger, R.D. (2004). The analysis of repeated measuremens with mixed-model adjusted *F* test. *Educational and Psichological Measurements,64*, 224-242.

Livacic-Rojas, P, Vallejo, Vallejo, & Fernández, (2010). Analysis of Type I Error Rates of Univariate and Multivariate Procedures in Repeated Measures Designs. *Communications in Statistics - Simulation and Computation*, *39: 3,* 624-640. DOI: 10.1080/03610910903548952

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of misspecifiying the error covariance structure in linear mixed models for longitudinal data, *Methodology, 4,* 10-21.

Vallejo, G., Arnau, J., Bono, R., Fernández, P., & Tuero, T. (2010). Selección de modelos anidados para datos longitudinales usando criterios de información y la estrategia de ajuste condicional. *Psicothema*, *22,* 323-333.

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Selecting the best unbalanced repeated measures model. *Behavior Research Methods, 43*, 18–36.

Wang, J. & Schaalje, G.B. (2009). Model selection for linear mixed models using predictive criteria. *Communication in Statistics-Simulation and Computation, 38,* 498-508.

ERRORES EN LA PRUEBA DE HIPÓTESIS DE LOS EFECTOS EN LOS DISEÑOS DE MEDIDAS REPETIDAS

M. P. Fernández[1], G. Vallejo[1], P. E. Livacic-Rojas[2], E. Tuero-Herrero[1], J. F. García[3] y M. Castillo-Fuentes[3]

[1] Universidad de Oviedo

[2] Universidad de Santiago de Chile

[3] Universidad de Valencia

Correo electrónico: paula@uniovi.es

Resumen

En esta investigación se examinan los errores cometidos en la prueba de las hipótesis subyacentes a dos modelos distintos satisfechos en los datos recogidos mediante un diseño de medidas repetidas. En concreto se examina el comportamiento de cinco estadísticos univariados. De ellos, sólo el AVAR asume una matriz de dispersión subyacente esférica. Los otros cuatro suponen que no lo es, pero de distinta manera. Dos procedimientos suponen que la correlación entre los datos no sigue un patrón particular y los otros dos suponen que existe autocorrelación serial de primer orden. El examen se realiza mediante métodos Monte Carlo. Se observó el comportamiento de los procedimientos bajo condiciones de normalidad en ausencia de esfericidad cuando existe autocorrelación serial de primer orden. También se examina el efecto de la falta de equilibrio entre el tamaño de los grupos, y la relación entre el tamaño del grupo y el tamaño de las matrices de dispersión. Los resultados nos indican que no sólo se debe de elegir cuidadosamente el estadístico de análisis en función de las características de los datos, sino que el investigador debe de anticipar qué modelo estadístico puede esperar para planificar cuidadosamente su investigación.

Los diseños de medidas repetidas, debido a que permiten sustraer las diferencias individuales del error experimental, son una estrategia eficaz para examinar el efecto de los tratamientos administrados en forma consecutiva, o estudiar la evolución de un comportamiento en el tiempo. La naturaleza de los niveles de la variable intrasujeto con frecuencia determina no sólo la naturaleza de la variable dependiente, sino también ciertas características que serán inherentes a ella. Así las cosas, es frecuente que si la variable intrasujeto son distintos niveles de una variable de tratamiento o tratamientos distintos, y son administrados aleatoriamente, los efectos de orden y/o residuales no supongan un problema, no solo sustantivo, sino estadístico tampoco. Sin embargo, si la variable dependiente es registrada en función de la edad o el tiempo, es esperable que la violación del supuesto de independencia

sea la opción más esperable. En esta situación es probable que se observen ciertas tendencias en relación con un proceso de maduración o de aprendizaje produciéndose efectos residuales y/o de autocorrelación capaces de dar lugar, además, a una cierta heterogeneidad en las varianzas. La importancia de lo anterior es crucial para decidir sobre la técnica estadística a utilizar para poner a prueba las diferentes hipótesis nulas del diseño. La consecuencia inmediata de violar el supuesto de esfericidad es que la hipótesis nula de los efectos del diseño son falsamente rechazada con más frecuencia de lo que deberían, y más aún cuanto mayor sea la desviación. Para resolver este problema, y en función de si o no los supuestos restantes del AVAR se cumplen, diferentes autores han desarrollado diferentes alternativas univariado, Greenhouse y Geisser (1959) (GG), Huynh y Feldt (1976) (HF) y Lecoutre (1991) (LEC). Sin embargo, todos asumen la correlación entre las puntuaciones de arbitraria. Debido a que es probable que exista dependencia entre las observaciones, algunos autores (Hearne, Clark y Hatch, 1983 y Jones, 1985, entre otros) han propuesto modelos univariado de la varianza que la tienen en cuenta. Hay abundante investigación sobre el comportamiento comportamiento de los procedimientos univariados que corregir la falta de esfericidad, sin asumir la existencia de autocorrelación serial. Sin embargo, existen muy pocos estudios sobre los procedimientos univariados que corregir la falta de esfericidad asumiendo que hay una correlación serial. Destacamos tres, Fernández, Vallejo, Livacic Rojas, Herrero y Cuesta (2008), Fernández, Vallejo, Livacic-Rojas, Herrero y Cuesta (2009) y Fernández, Vallejo y Livacic Rojas (2010). Los anteriores autores examinaron el comportamiento con respecto a error de tipo I (para el efecto principal intrasujeto y la interacción) y la potencia (para el efecto principal intra-sujeto) de seis estimadores univariados, el AVAR, las pruebas GG, HF, de Lecoutre (1991) (LEC), de la Hearne et al, (1983) (HCH), y de Jones (1985) (JN), en un diseño factorial mixto (p x q). Sin embargo, la potencia de prueba de estos métodos conjuntamente nunca se ha examinado para estudiar el efecto de la interacción.

El objetivo de esta investigación ha sido analizar el comportamiento de los seis estadísticos univariados referidos anteriormente con respecto a la potencia de prueba de la interacción en un diseño factorial mixto (3x4).

MÉTODO

Llevamos a cabo un estudio de simulación Montecarlo para un diseño Split-Plot de medidas repetidas (3x*q)*. Examinamos los procedimienros AVAR, GG, HF, LEC, HCH y JN bajo distribución normal multivariada. Manipulamos las siguientes variables:

1. *Tamaño de muestra total*. Utilizamos tres tamaños totales de muestra (N): N=15, 30 y 48. El valor del coeficiente de variación muestral fue 0 cuando el diseño fue equilibrado (n*j*=5, 10 y 16 respectivamente en los tamaños N anteriores), y 0.33 cuando fue no equilibrado (N= 15, los tamaños grupales fueron: n*j*=3, 5 y 7; N= 30, n*j*=6, 10 y 14; N= 48, n*j*=9, 16 y 23).

2 *Ocasiones de medida:* (*q*): 4, 6, 8 y 12.

3. *Patrones de covarianza empleados para generar los datos*: Los datos fueron generados utilizando tres estructuras de covarianza autorregresivas. Las matrices (AR[1]) manifiestan estacionariedad en las varianzas y la correlación entre la *k*th y la *k*'th observación es ρ| k-k'|, ρ= [-0.8:0.8: (0.2)]. Las matrices ARH[1] expresan matrices de covarianza con el mismo diseño de correlación serial positiva y negativa que las matrices AR[1] pero las varianzas son heterogéneas (forma estructurada). De éstas contemplamos dos condiciones de no estacionariedad: creciente AREH[1]-C y decreciente AREH[1]-D . Las matrices AR[1] y AREH[1] tienen una desviación de la esfericidad que puede ser calculada mediante ρ y cuyo tamaño está en función de cada uno de los elementos utilizados en su construcción.

4. *Igualdad de las matrices de dispersión*. Los elementos de las tres matrices mantenían entre sí las relaciones siguientes: Σ1=1/3Σ2 y Σ3=5/3Σ2.

5. *Emparejamiento de las matrices de covarianza y el tamaño de los grupos*. Razón C/H. H0 (diseño está equilibrado, heterogeneidad entre grupos), H+ (diseño no balanceado, el grupo de menor tamaño se asocia con la matriz de dispersión menor) y H- (diseño no balanceado, el grupo de menor tamaño se asocia con la matriz de dispersión mayor)

6. Se proponen dos situaciones hipotéticas con las que un investigador aplicado se puede encontrar y en las que subyace un modelo no aditivo, donde, naturalmente, la interacción es el efecto interesante del modelo. Ambas suponen una interacción no ordinal. En la primera las medias de los niveles de la variable entre grupos son distintas y las medidas en los niveles de la variable intra-sujetos experimentan tendencias también distintas. Se examinan dos situaciones: tamaño del efecto pequeño y tamaño del efecto mayor). En la segunda situación las medias en los niveles de la variable entre grupos son iguales para los de los grupos y las medias en los niveles de la variable intra-sujetos experimentan dos tendencias distintas. Se examinan las mismas dos situaciones que en el primer caso, pero en esta ocasión el tamaño del efecto es menor.

RESULTADOS

En la Tabla 1 se muestra los resultados para un subconjunto seleccionado de las condiciones estudiadas que muestran las diferencias entre los distintos procedimientos. En la Tabla 1 se resalta la potencia de prueba empírica que es más elevada que la potencia teórica. se tomó como referencia el valor medio de la Potencia de Prueba Teórica calculada sobre el procedimiento de Greenhause y Geisser para nj =5 y q =4 para todos los magnitudes de correlación sometidas a estudio en esta investigación. La potencia de prueba teórica cuando las matrices son AR[1], ARH[1]-I, ARH[1]-D: (ρ + = 0.77; ρ - = 0.68).

Tabla 1. Potencia de Prueba empirica para la FV Interacción. Estructuras de Covarianza: AR[1], ARH[1]-I y ARH[1]-D, Correlación Positiva y Negativa. (α=.05). Distancia entre las medidas pequeña

			Situación 1 ; N=30												Situación 2 ; N=48											
			Correlation serial positiva						Correlation serial negativa						Correlation serial positiva						Correlation serial negativa					
		ρ	AVAR	GG	HF	LEC	HCH	JN	AVAR	GG	HF	LEC	HCH	JN	AVAR	GG	HF	LEC	HCH	JN	AVAR	GG	HF	LEC	HCH	JN
AR[1] f_{AR+}=0.309 f_{AR-}=0.219	B H0	.20	.622	.584	.624	.609	.614	.555	.561	.518	.564	.547	.548	.618	.856	.843	.854	.853	.852	.805	.810	.792	.812	.804	.801	.885
		.40	.637	.588	.627	.611	.615	.510	.449	.379	.422	.405	.402	.701	.859	.837	.815	.846	.847	.760	.732	.676	.702	.693	.689	.923
		.60	.622	.557	.594	.578	.580	.453	.321	.229	.261	.246	.239	.755	.848	.813	.827	.821	.819	.702	.533	.413	.439	.427	.421	.946
		.80	.619	.540	.572	.557	.557	.413	.198	.114	.126	.120	.116	.799	.841	.792	.804	.799	.799	.664	.277	.154	.166	.158	.155	.967
	B H+	.20	.411	.376	.422	.405	.401	.077	.316	.281	.325	.309	.306	.124	.671	.648	.675	.664	.663	.172	.573	.543	.576	.563	.562	.282
		.40	.436	.389	.430	.414	.411	.060	.229	.183	.217	.201	.200	.148	.683	.648	.671	.663	.663	.134	.449	.385	.418	.403	.400	.360
		.60	.441	.375	.414	.398	.394	.042	.153	.100	.118	.110	.105	.192	.683	.630	.651	.642	.642	.098	.270	.184	.203	.193	.189	.446
		.80	.446	.364	.397	.382	.373	.035	.105	.054	.061	.056	.054	.236	.672	.594	.616	.605	.604	.074	.131	.071	.077	.073	.072	.519
	B H-	.20	.721	.676	.709	.696	.715	.872	.678	.627	.666	.650	.666	.917	.894	.876	.882	.885	.891	.968	.877	.856	.867	.862	.870	.983
		.40	.718	.670	.669	.687	.700	.847	.612	.526	.569	.549	.563	.937	.893	.873	.882	.878	.884	.954	.817	.768	.787	.779	.784	.987
		.60	.723	.666	.690	.678	.692	.837	.477	.362	.400	.382	.382	.946	.882	.850	.860	.856	.863	.941	.679	.567	.593	.581	.580	.993
		.80	.707	.635	.660	.647	.650	.811	.314	.197	.215	.203	.200	.962	.866	.824	.834	.829	.832	.992	.424	.282	.295	.278	.283	.996
ARH[1]-D f_{ARH-D+}=0.301 f_{ARH-D-}=0.214	B H0	.20	.409	.375	.419	.400	.402	.351	.354	.317	.362	.345	.342	.448	.650	.625	.649	.640	.643	.577	.570	.540	.569	.558	.558	.678
		.40	.434	.390	.427	.412	.411	.324	.282	.227	.264	.248	.247	.491	.649	.615	.636	.627	.631	.521	.477	.413	.445	.431	.429	.735
		.60	.437	.376	.412	.396	.393	.292	.201	.137	.157	.147	.143	.521	.644	.593	.613	.605	.606	.478	.325	.234	.252	.243	.242	.781
		.80	.440	.362	.394	.379	.375	.263	.157	.088	.099	.092	.090	.571	.628	.561	.578	.570	.570	.416	.187	.103	.111	.105	.105	.810
ARH[1]-I f_{ARH-I+}=0.602 f_{ARH-I-}=0.440	B H0	.20	.979	.970	.977	.975	.978	.965	.974	.956	.967	.963	.970	.991	.999	.999	.999	.999	.999	.999	.999	.999	.999	.999	.999	.999
		.40	.975	.962	.969	.971	.971	.942	.943	.894	.917	.908	.921	.992	.999	.999	.999	.999	.999	.997	.998	.996	.997	.996	.997	1
		.60	.964	.945	.954	.950	.956	.909	.824	.675	.718	.693	.722	.996	.998	.997	.997	.997	.998	.992	.989	.963	.968	.965	.973	1

Nota: NBH0= Distribución normal, diseño balanceado y heterogeneidad entre los grupos; NnBH+= Distribución normal, diseño no balanceado y heterogeneidad entre los grupos con relación positiva; NnBH-= normal, diseño no balanceado y heterogeneidad entre los grupos relación negativa.

1. La potencia empírica incrementa conforme incrementa *f*, n*j* y ε. La evolución de esta conducta para estos parámetros se observa muy claramente cuanto todos ellos tienen pequeños valores. Para valores elevados de *f*, n*j* y ε las diferencias desaparecen.
2. Bajo correlación negativa, un incremento de la misma causa una reducción en la potencia en cada matriz Σ estudiada. Cuando la correlación es positiva y la matriz Σ es AR [1], ARH [1]-D y ARH [1]-I la potencia empvrica permanece casi estable a lo largo de todos los valores de la correlaciσn, pero cuando la matriz Σ es ARAH [1], un incremento en el tamaño de la correlación positiva siempre implica un incremento de la potencia empírica en todos los procedimientos.
4. Cuando el tamaño de los grupos y el tamaño de las matrices de dispersión están directamente apareados la potencia empírica fue siempre menor. Cuando el tamaño de los grupos y el tamaño de las matrices de dispersión están inversamente apareados la potencia empírica fue siempre más elevada. El procedimiento JN fue el más sensible a estas condiciones.

CONCLUSIONES

Los procedimientos AVAR, GG y HF han sido ampliamente estudiados y los resultados hallados están en sintonía con los hallados por otros autores. Los procedimientos AVAR, JN, a pesar de que muestran una elevada potencia empírica en muchas situaciones nunca deben ser procedimientos elegidos para analizar los datos debido a su escasa protección con respecto al error de Tipo I (Fernández et al., 2008, 2009, 2010 a, 2010b). Los procedimientos LEC y HCH que han demostrado su robustez (Fernández et al., 2008, 2009, 2010 a, 2010 b), por eso, y teniendo en cuenta los resultados hallados aquí, sería cualquiera de ellos una elección apropiada.

Recomendación: El investigador debe planificar el tamaño de la muestra necesario para poner a prueba los efectos deseados, y también, estimar, anticiparse al modelo estadístico que puede subyacer en el diseño recogiendo sus datos, porque la misma distancia entre las medias sobre el tiempo no siempre logra un mismo tamaño del efecto para la interacción en cualquier modelo estadístico.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado mediante el proyecto de investigación concedido por el MCI (Ref.: PSI-2008-03624).

REFERENCIAS

Fernández, P.; Vallejo, J.; Livacic-Rojas; Herrero, J. & Cuesta, M. (2010). Comparative robustness of six tests in repeated measures designs with specified de-

partures from sphericity. *Quality & Quantity, Quality & Quantity, 44*(2), 289-301. DOI: 10.1007/S11135-008-9198-3.

Fernández, P., Vallejo, G., Livacic-Rojas, P., Herrero, J. y Cuesta, M. (2009). Comparison of the power of three statistics in repeated measures design in the absence of sphericity with and without serial autocorrelation. Review of Psychology, 16(2), 65- 76.

Fernández, P.; Vallejo, G. & Livacic-Rojas, P. (2008, July). Comparison of the robustness of the SPSS MIXED procedure with regard to another three univariate statistics in repeated measures designs with specified departures from sphericity. *Paper presented at the III European Congress of Methodology*, Oviedo (Spain).

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Hearne, E.M., Clark, G.M. & Hatch, J.P. (1983). A test for serial correlation in univariate repeated-measures analysis. *Biometrics, 39*, 237-243.

Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data randomized block and split-plot designs. *Journal of Educational Statistics, 1,* 69-82.

Jones, R.H. (1985). Repeated measures, interventions, and time series analysis. *Psychoneuroendocrinology, 10*(1), 5-14.

Lecoutre, B. (1991). A correction for the $\widetilde{\varepsilon}$ approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics, 16*, 371-372.

INVESTIGACIÓN CUASI-EXPERIMENTAL Y PRE-EXPERIMENTAL PUBLICADA EN ESPAÑA EN LA ÚLTIMA DÉCADA

E. Tuero-Herrero[1], M. P. Fernández[1], G. Vallejo[1] y P. Livacic-Rojas[2]
[1] Universidad de Oviedo
[2] Universidad de Santiago de Chile
Correo electrónico: tuero.eli@gmail.com

Resumen

Realizamos un estudio descriptivo mediante análisis de documentos con el objetivo de conocer cuáles son los diseños cuasi-experimentales y pre-experimenlales más utilizados en las investigaciones publicaciones en revistas científicas de Psicología editadas en castellano, y sus principales características. El período temporal estudiado ha sido desde 1999 a 2009. Elegimos tres revistas científicas españolas de Psicología, a saber, Psicothema, Internacional Journal and Health Psychology (IJCHP) y Psicológica. Las tres cumplían determinados objetivos que consideramos importantes. La unidad de análisis ha sido cada uno de los estudios cuasi-experimentales y pre-experimentales publicados en un mismo artículo. Hemos encontrado que el diseños más elegido en las investigaciones cuasi-experimentales es el diseño de grupo control no equivalente. El análisis univariado de la varianza (t y F) es la técnica de análisis de datos preferida. También hemos observado que el reporte de resultados estadístico es deficiente en la mayoría de los casos.

El término Cuasi-Experimental fue propuesto por Campbell y Stanley (1966) para referirse a las características peculiares que adquiere una investigación (en la que se manipulaba al menos una variable independiente) amparada en un enfoque hipotético-deductivo (con finalidad explicativa causal) en contextos sociales como resultado del control insuficiente en la asignación de los sujetos a los grupos por razones diversas (restricciones institucionales, presiones políticas, consideraciones económicas o éticas). Otras veces la participación en un cuasiexperimento está determinada por un proceso de autoselección –voluntarios–, o la exposición a los tratamientos viene determinada por su puntuación en la V.D. En cualquiera de estos casos el control de variables extrañas conocidas es imposible mediante técnicas de control experimental, y el control de variables extrañas no conocidas es impracticable debido a la ausencia de la asignación aleatoria de las unidades experimentales a las condiciones de la investigación.

Fue a raíz de esta preocupación por preservar la validez de la inferencia causal, que Cook y Campbell (1979), además de ampliar el concepto de validez, le dieron una condición importante al término *Cuasi-experimental*, y lejos de ser un mero calificativo para aquellas investigaciones que implicaban comparación formal pero no discurrían en óptimas condiciones, las otorgaban un estatus propio constituyendo una nueva metodología de investigación con ánimo de garantizar inferencias causales a través del desarrollo de un conjunto de técnicas de diseño y análisis estadístico útiles en investigaciones de naturaleza social aplicada, donde sí existe manipulación de la variable independiente, pero la aleatorización y el control directo y riguroso no son posibles. A partir de entonces y hasta hoy, una cascada de metodólogos se ha ocupado de estos diseños tanto a nivel metodológico como analítico.

Los diseños Cuasi-Experimentales parten todos ellos de tres modelos básicos, que nosotros denominaremos Diseños Pre-Experimentales debido a que ninguno de ellos tiene una adecuada condición de comparación que permita extraer conclusiones válidas de un modo digno, a saber, diseño de un solo grupo con medida post, diseño de un solo grupo con medidas pre-post y diseño con dos o más grupos sólo con medidas post (en todos los casos, el/los grupo(s) no son formados aleatoriamente). Todos los Diseños Cuasi-Experimentales se construyen sobre estos tres modelos básicos añadiendo medidas, añadiendo grupos o añadiendo ambos. Consideramos cuatro grandes conjuntos de Diseños Cuasi-Experimentales: diseños Cuasi-Experimentales de un solo grupo, de Grupo Control No Equivalente, de Discontinuidad en la regresión y de Series temporales.

Realizamos un estudio descriptivo mediante análisis de documentos con el objetivo de conocer cuáles son los diseños cuasi-experimentales y pre-experimenlales, (en adelante D.Cx. y D.Px. respectivamente) más utilizados en las investigaciones publicaciones en revistas científicas de Psicología editadas en castellano, y sus principales características. En este reporte se expone una pequeña parte de una investigación que sigue en curso.

MÉTODO

Materiales

Se eligieron arbitrariamente tres revistas científicas de psicología editadas en España que cumplían tres criterios: a) ser revistas de temática general en el campo de la Psicología; b) aparecer en el *Journal Citation Reports (JCR)* con factor de impacto (en 2009); c) aparecer en IN-*RECS* con índice de impacto (en 2009) dentro del primer cuartil. Las revistas fueron Psicothema, Internacional Journal and Health Psychology (IJCHP) y Psicológica.

Período temporal revisado: Re han revisado 11 años, los comprendidos en el período 1999-2009. Se han exceptuado cuatro suplementos de la revista Psicothema (2 en el año 2000, 1 en 2002 y otro más en 2006), sin embargo no se excluyeron

los denominados números especiales y monográficos en la revista Psicológica porque suponen siempre un número habitual de la revista. Se han revisado un total de 1685 artículos científicos.

Unidad de análisis: La unidad de análisis ha sido el estudio, considerando unidades independientes cada uno de los estudios Cuasi-Experimentales y Pre-Experimentales publicados en un mismo artículo.

Diseño y Procedimiento: Dos grupos de expertos (profesores de Psicología) compuestos por dos profesionales cada grupo analizaron aleatoriamente la mitad de los artículos. En cada grupo cada artículo era analizado por cada experto en las variables de interés. Posteriormente se comparaban las clasificaciones obtenidas por ambos. El porcentaje medio de acuerdo entre los grupos fue de 95%. En aquellos casos en los que no coincidían se revisaban otra vez con la colaboración de un tercer experto y se analizaba el contenido del artículo hasta que los tres expertos llegasen a una conclusión compartida.

Variables registradas: Se ha examinado la *prevalencia*, tipo de diseño, modo de elección del tamaño de la muestra, número de variables dependientes, modo de analizar los datos de las investigaciones y exposición de los resultados..

RESULTADOS

Únicamente hemos encontrado 66 investigaciones Cuasi y Pre experimentales que constituyen el 3.91% de las investigaciones publicadas en estas tres revistas. La revista Psicothema publica el 71.2% de estas investigaciones (n=47), seguida por la revista IJCHP que publica el 22.5% (n=15). El restante 6.1% lo publica la revista Psicológica (n=4). No existe una tendencia a lo largo de los años en cuanto a la cantidad de estas investigaciones en las revistas Psicothema y Psicológica (podemos calificar que la presencia de investigaciones Cuasi-Experimentales en esta revista es anecdótica), pero sí parece que existe en la revista IJCHP, que en sus 5 primeros años (2000-2004) sólo publicó dos de éstas investigaciones, y en lo 5 años siguientes publicó 13. Tampoco existe ninguna tendencia a lo largo de estos años en lo que al tipo concreto de diseño respecta (ver Tabla 1).

Los Diseños Cuasi-Experimentales son más abundantes (n=41; 62,12%) que los Diseños Pre-Experimentales (n=25; 37.87%). En lo que respecta a los D.Cx., el 80.48% son diseños de grupo control no equivalente (de diversa formación) y el 17.07% son Diseños de Discontinuidad en la regresión. Diseños Cuasi-experimentales de un solo grupo sólo existe uno. No hay ningún diseño de Cohortes ni de Series Temporales. En lo que respecta a los D.Px. los más abundantes son los diseños de un solo grupo con medida pre y post (44%) (ver Tabla 1).

En el 97% de todas las investigaciones (D.Cx. y D. Px.) no se estima a priori el tamaño de la muestra. Únicamente se lleva a cabo esta tarea en dos investigaciones Px., y en una de ellas, se estima una vez que ya se tiene la muestra con ánimo de saber si fue poca o mucha (fue suficiente en este caso).

Tabla 1. Diseños de Investigación Cuasi-Experimentales y Pre-Experimentales llevados a cabo en investigaciones empíricas publicadas en tres Revistas científicas de Psicología, IJCHP, Psicológica y Psicothema durante el período temporal de 11 años (1999-2009)

							D. Cuasi-Experimentales			D. Pre-Experimentales		
		IJCHP	**Psicológica**	**Psicothema**	**Total n**	**p**	**D.Cx-1G**	**D.Cx-GCNE**	**D.Cx-DR**	**D.Px-1G**	**D.Px-Gs**	**Total**
Año	2000	0	1	4	5	7,6%		3	1		1	
	2001	0	0	1	1	1,5%					1	
	2002	0	0	8	8	12,1%		6	1		1	
	2003	1	0	4	5	7,6%		1	1	2	1	
	2004	1	2	7	10	15,2%	1	5	1	1	2	
	2005	3	1	1	5	7,6%		1	2	2		
	2006	2	0	6	8	12,1%		4		3	1	
	2007	4	0	5	9	13,6%		7		1	1	
	2008	2	0	3	5	7,6%		4			1	
	2009	2	0	8	10	15,2%		2	1	4	3	
Total	n	15	4	47	66	100.00	1	33	7	13	12	66
	p	22.7%	6.1%	71.2%	100%	100.00	1,5%	50%	9%	19,2%	18%	100%
							Tipo Concreto					
Diseño	D.Cx-1G	0	0	1	1	1,5%						
	D.Cx-GCNE	5		28	33	50%						
	D.Cx-DR	2	1	4	7	9%						
	D.Px-1G	5		8	13	19,2%						
	D.Px-Gs	3	3	6	12	18,2%						
Total	n	15	4	47	66	100%						
	p	22.7%	6.1%	71.2%	100%							

Leyenda: D.Cx-1G= Diseño Cuasi-Experimental de un solo grupo; D.Cx-GCNE= Diseño Cuasi-Experimental de Grupo Control no equivalente; D. Cx-DR= Diseño Cuasi-Experimetal de Discontinuidad en la Regresión; D.Px-1G= Diseño Pre-Experimental de un solo Grupo; D.Px-Gs= Diseño Pre-Experimetal de más de un grupo; n= Frecuencia; p= porcentaje.

La opción más elegida para analizar los datos en los diseños cuasi-experimentales es el análisis univariado paramétrico (mediante AVAR o *t* de Student). Por ejemplo, en los D.Cx. clásicos de dos grupos (experimental y control) y dos medidas (pre-post), se lleva a cabo en 18 investigaciones (94.73%). Sólo en 2 investigaciones (10%) se realiza un análisis no paramétrico. El análisis multivariado sólo se utiliza en 3 investigaciones (15.8%) y el análisis de la covarianza se realiza en 5 investigaciones. En los D. Px. no hay nada nuevo que destacar, salvo que es en estos diseños en los que el análisis de la regresión se utiliza en más ocasiones que en los D.Cx. .

Un porcentaje muy elevado de las investigaciones expone el valor empírico del estadístico, los gl y p. Con respecto a ésta última, sólo el 62.1% exponen su valor exacto, el resto se limitan a escribir > que, o < que α. El Tamaño del efecto no se aporta en el 80.3% de las investigaciones. La exposición de la potencia de prueba empírica y de los intervalos confidenciales brillan por su ausencia. Sólo en el 19,7% (n=13) de las investigaciones se expone qué paquete estadístico utilizaron para analizar los datos, y en todas las que lo exponen utilizaron el SPSS.

CONCLUSIONES

No cabe duda de que los estudios orientados a la evaluación de las publicaciones realizadas en Revistas Científicas desde distintos puntos de vista cotiza y ha cotizado al alza. Son muchos los estudios realizados sobre la investigación experimental (Fernández, Vallejo, Livacic-Rojas y Tuero-Herrero (en preparación)) desde diferentes ópticas (análisis llevados a cabo, tipo de diseño, examen de resultados, temática científica…), sin embargo, y en relación a la investigación experimental, los orientados a estudiar la investigación Cuasi-experimental son más bien escasos.

Hemos encontrado que la investigación cuasi-experimental es muy inferior en número a la experimental. En otros estudios realizados sobre otras revistas de Psicología editadas en castellano, y sobre publicaciones realizadas en la década de 1990, Sánchez Meca, Valera, Velandrino y Marín (1992), Valera, Sánchez, Marín y Velandrino (1998) y Valera, Sánchez Meca y Marín (2000), han hallado que es la investigación cuasi-experimental más abundante que la experimental. Scandura y Williams (2000) obtienen resultados similares en revistas orientadas a la Psicología de las organizaciones. En el ámbito de la educación Gersten, Baker, Smith-Johnson, Flojo, y Hagan Burke (2004), Pei-Hsuan Hsieh, et al (2005), Seethaler and Fuchs (2005) y Whitehurst (2003) han hallado que en las revistas de Psicología de la educación abundaban más las investigaciones experimentales que las cuasi-experimentales, y que en revistas de otros ámbitos de la educación había mayor proporción de investigaciones cuasi-experimentales. Quizás pase esto aquí

Hemos encontrado que los D. Cx son más abundantes que los D. Px. Harris, Bradham, Baumgarten, Zuckerman, Fink and Perencevich, (2004), Harris, A.D.; Lautenbach, E. and Perencevich, E. (2005) y Harris, McGregor, Perencevich, et al.

(2006) en el ámbito de las enfermedades infeccionas hallaron justo lo contrario. Sin embargo, sí que hubo una coincidencia, y es que los D. Cx. más habituales fueron de grupo control no equivalente clasicazos (2x2), y de los D.Px, los diseños de un solo grupo pre-port.

Analizar múltiples variables dependientes también es una coincidencia con Scandura y Williams (2000) y Pei-Hsuan Hsieh, et al (2005). El Análisis de la varianza también es la más utilizada en consenso con todas las investigaciones anteriores.

Se puede apreciar en dicha Tabla que en un porcentaje muy elevado se expone el valor empírico del estadístico, los gl y p. Con respecto a ésta última, sólo el 62.1% exponen su valor exacto, el resto se limitan a escribir > que, o < que α. El Tamaño del efecto sólo se aporta en el 80.3% de las investigaciones, la exposición de la potencia de prueba empírica y de los intervalos confidenciales brillan por su ausencia. Sólo en el 19,7% (n=13) de las investigaciones se expone qué paquete estadístico utilizaron para analizar los datos, y en todas las que lo exponen utilizaron el SPSS.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado mediante el proyecto de investigación concedido por el MCI (Ref.: PSI-2008-03624).

REFERENCIAS

Fernández, M.P., Vallejo, G., Livacic-Rojas, P. y Tuero-Herrero, E. (en preparación). El diseño experimental en tres revistas científicas de Psicología en más de una década.

Sánchez, J., Valera, A., Velandrino, A.P. y Marín, F. (1992). Un estudio de la potencia estadística en Anales de Psicología. *Anales de Psicología, 8*, 19-32.

Valera, A., Sánchez Meca, J. y Marín, F. (2000). Contraste de hipótesis e investigación psicológica española: análisis y propuestas. *Psicothema 12*(2), 549-552

Valera, A., Sánchez, J., Marín, F. y Velandrino, A.P. (1998). Potencia Estadística de la Revista de Psicología General y Aplicada (1990-1992). *Revista de Psicología General y Aplicada. 51*(2).

Gersten, R., Baker, S. K., Smith-Johnson, J., Flojo, J. R., & Hagan-Burke, S. (2004). A tale of two decades: Trends in support of federally funded experimental research in special education. *Exceptional Children, 70,* 323–332

Harris, A.D., Bradham, D.D., Baumgarten, M., Zuckerman, I.H., Fink, J.C., & Perencevich, E.N. (2004). The Use and Interpretation of Quasi-Experimental Studies in Infectious Diseases. *Clinical Infectious Diseases,38*, 1588-91

Harris, A.D., Lautenbach, E., & Perencevich, E. (2005). A systematic review of quasi-experimental study designs in the fields of infection control and antibiotic resitance. *Clinical Infectious Diseases*. 41(1), 77–82

Harris, A.D., McGregor, J.D., Perencevich, E.N., Furuno, J.P., Zhu, J., Peterson, D.E. & Finkelstein, J. (2006). The Use and Interpretation of Quasi-Experimental Studies in Medical Informatics. *Journal of the American Medical Informatics Association, 13* (1), 16-23.

Peggy (Pei-Hsuan) Hsieh, P., Acee, T., Chung, WH., Hsieh, YP., Kim, H., Thomas, G.D., You, Ji., Levin, J.R. & Robinson, D.H.. (2005). Is Educational Intervention Research on the Decline?. *Journal of Educational Psychology, 97*(4), 523–529

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends and implications for future research. *Academy of Management Journal, 43*, 1248-1264.

Seethaler, P. M., & Fuchs, L. S. (2005). A drop in the bucket: Randomized controlled trials testing reading and math interventions. *Learning Disabilities Research & Practice, 20,* 98–102.

Whitehurst, G. J. (2003, April). *The Institute of Education Sciences: New wine, new bottles.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

LA SIMULACIÓN COMO INSTRUMENTO DE INVESTIGACIÓN: MÉTODOS Y APLICACIONES

Coordinador: Jaume Arnau
Universidad de Barcelona

En los últimos años, se ha producido en nuestro país una emergencia de estudios basados en procedimientos de simulación. La prueba la tenemos en la cantidad de publicaciones que van apareciendo en revistas de carácter internacional y de impacto. En línea con ello, el propósito de este simposio, es dar cabida a las diferentes líneas de trabajo que están aplicando la simulación como instrumento científico y ofrecer la posibilidad de dar a conocer cuáles son los modelos, programas y objetivos que persiguen los diferentes grupos de trabajo diseminados a lo largo de nuestra geografía. Se pretende, en primer lugar, cotejar los diferentes programas de simulación que estos distintos grupos manejan y poner en común las ventajas relativas de los procedimientos aplicados, así como de sus posibilidades de cara a la implementación de los modelos tantos teóricos como estadísticos. En segundo lugar, conocer para que y en qué sentido se usa el método de trabajo basado en la simulación. ¿Qué aporta la simulación al conocimiento de los procesos teóricos y estadísticos? ¿Sirve para indagar las propiedades de los modelos estadísticos cuándo se trabaja con datos de distribución normal y no normal? ¿Permite descubrir propiedades desconocidas sobre el comportamiento de los ajustes estadísticos? Por último, es interesante dar a conocer los ámbitos de estudio en los que se está trabajando con esta clase de procedimientos. Si los programas que se utilizan para simular poseen características comunes o no, y la razón por la que son utilizados. A su vez, si la simulación es requerida por el contexto en que se trabaja, por la tradición de la línea en que el grupo de investigación está instalado, por la facilidad de definir los modelos o por las facilidades que ofrece el programa. Cabría plantearnos también el uso de las macros dentro de la simulación y si realmente adquieren un papel relevante en los distintos procedimientos de simulación que se han presentado. En suma, se trata de un Simposio que sirva para canalizar las distintas corrientes basadas en la simulación y que se convierta en una plataforma para conocer de cerca los distintos grupos de trabajos y sus herramientas particulares de investigación dentro de nuestro país.

PALABRAS CLAVE: Simulación, Robustez, No-normalidad, Medidas repetidas, Modelos multinivel, Diseños de caso único.

ESTUDIO DE LA ROBUSTEZ CON DATOS DE DISTRIBUCIÓN LOGNORMAL. DATOS DE MEDIDAS REPETIDAS

Jaume Arnau y Roser Bono
Universidad de Barcelona
Correo electrónico: rbono@ub.edu

Resumen

Dado que la mayoría de distribuciones en investigaciones del ámbito de las ciencias sociales y de la salud se apartan de la normalidad, se ha llevado a cabo un estudio de simulación con datos de distribuciones más cercanas al mundo real como, por ejemplo, la distribución lognormal. En primer lugar, se generaron los datos de distribución lognormal, mediante el sistema SAS, siguiendo un diseño split-plot. En segundo lugar, se generaron las matrices de covarianza no-estructurada de la población para diferentes cantidades de medidas repetidas (K = 4, 6 y 8) y valores de esfericidad de (ε = 0,57 y 0,75). En tercer lugar, con el PROC MIXED del SAS, se selecciona la estructura de covarianza de mejor ajuste según el criterio Akaike, y se analizan los distintos conjuntos de datos con los grados de libertad corregidos por el procedimiento Kenward-Roger (KR). De los resultados obtenidos, se concluye que las estimaciones del efecto tiempo son menos robustas con distribuciones lognormales que con distribuciones normales, especialmente con grupos balanceados y emparejamiento nulo. En cuanto a las estimaciones del efecto de interacción, se observa un efecto contrario, es decir, la prueba tiende a ser más robusta con distribución lognormal.

Los datos de las simulaciones se han generado mediante una serie de macros creadas con el SAS 9.2 (SAS Institute, 2008). En primer lugar, se generaron las matrices de covarianza a partir de varianzas y correlaciones con valores de esfericidad $\varepsilon = 0.57$ y $\varepsilon = 0.75$. Posteriormente, con el generador *rannor* se obtuvieron observaciones pseudoaleatorias de distribución normal mediante el factor Cholesky de la matriz de covarianza $\boldsymbol{\Sigma}_i$. Los datos de distribuciones lognormales se generaron con el mismo procedimiento pero transformados por los coeficientes de Fleishman (1978)

En la Tabla 1 se muestran las distintas combinaciones de las variables examinadas en este estudio. De cada combinación se realizaron 1000 réplicas a un nivel de significación de 0.05, tanto para distribuciones normales como no-normales. Para generar los datos, se usó la estructura de covarianza UN, por ser la más típica con datos longitudinales.

Tabla 1. Tamaños de grupo para diseños split-plot con J = 3 y K = 4, 6, 8

N	N_1	n_2	n_3	Δn_j	Covarianza entre-grupos	Emparejamiento
30	10	10	10	0	=	Nulo
36	12	12	12	0	=	Nulo
42	14	14	14	0	=	Nulo
30	10	10	10	0	≠	Nulo
36	12	12	12	0	≠	Nulo
42	14	14	14	0	≠	Nulo
30	5	10	15	0.41	≠	+
36	6	12	18	0.41	≠	+
42	7	14	21	0.41	≠	+
30	15	10	5	0.41	≠	–
36	18	12	6	0.41	≠	–
42	21	14	7	0.41	≠	–

Note: J: grupos; *K*: número de medidas repetidas; *N*: tamaño de muestra total; n_1, n_2 y n_3: tamaños de grupo; Δn_j: coeficiente de varianza del tamaño de grupo; =/≠: homogeneidad/heterogeneidad de matrices de covarianza entre-grupos; +/-: emparejamiento positivo/negativo de tamaños de grupo y matrices de covarianza.

RESULTADOS

En las Tablas 2 y 3 se recogen los porcentajes de ajuste de las matrices de covarianza más comunes a la matriz de población UN, para covarianzas entre-grupos homogéneas y heterogéneas, respectivamente.

En la Tabla 2, se observa que con distribución normal en un 66.7% la estructura que mejor se ajusta es la de población UN. Con ε = 0.57, se ajusta bien en todos los casos. En cambio, cuando ε = 0.75 se ajusta mejor sólo para $K = 4$ y al aumentar la cantidad de medidas repetidas la matriz CSH tiene mejor ajuste, 33.3%. Respecto a la distribución lognormal, en ningún caso se ajusta mejor la matriz de población, sino que la de mejor ajuste es la matriz UN_j en un 100%.

Al analizar la Tabla 3 teniendo en cuenta el tipo de emparejamiento, se observa que con distribuciones normales, sólo cuando el emparejamiento es nulo la estructura de covarianza que mejor se ajusta coincide con la estructura de población UN_j en un 22.2%. Con la distribución lognormal, en un 33.4% se ajuste mejor la matriz de población UN_j para cualquier valor de K. Con emparejamientos positivos y negativos, se ajusta mejor la matriz de población UN_j sólo cuando $K = 4$, tanto con distribuciones normales como lognormales. Al aumentar la cantidad de medidas repetidas, la matriz ARH_j tiene un mejor ajuste en todas las distribuciones, especialmente cuando ε = 0.75. En cambio, cuando ε = 0.57 y $K = 6$, en todas las distribuciones se ajusta bien la matriz UN.

Tabla 2. Porcentajes de ajuste de las estructuras de covarianza más comunes (UN, CSH, ARH y UNj) a la matriz de covarianza de la población UN. Covarianzas entre-grupos homogéneas

	Matrices de covarianza ajustadas					
	UN		**CSH**	**ARH**	**UN$_j$**	
Distribuciones	ε = 0.57	ε = 0.75	ε = 0.75	ε = 0.57	ε = 0.57	ε = 0.75
Normal	*K* = 4, 6, 8	*K* = 4	*K* = 6, 8			
	50	16.7	33.3			
Totales	66.7					
Lognormal					*K* = 4, 6, 8	*K* = 4, 6, 8
(γ_1=1.75 γ_2=5.9)					50	50
Totales					100	

Tabla 3. Porcentajes de ajuste de las estructuras de covarianza más comunes (UN, UN$_j$ y ARH$_j$) a la matriz de la población UN$_j$. Covarianzas entre-grupos heterogéneas

	Emparejamiento					
	Nulo		**+**		**-**	
	ε=0.57	**ε=0.75**	**ε=0.57**	**ε=0.75**	**ε=0.57**	**ε=0.75**
Distribuciones	**UN**					
Normal			*K*=6		*K*=6	
			5.6		5.6	
Lognormal			*K*=6		*K*=6	
			5.6		5.6	
	UN$_j$					
Normal	*K*=6,8	*K*=4,8	*K*=4	*K*=4	*K*=4	*K*=4
	11.1	11.1	5.6	5.6	5.6	5.6
Totales	22.2		11.2		11.2	
Lognormal	*K*=4,6,8	*K*=4,6,8	*K*=4	*K*=4	*K*=4	*K*=4
	16.7	16.7	5.6	5.6	5.6	5.6
Totales	33.4		11.2		11.2	
	ARH$_j$					
Normal	*K*=4	*K*=6	*K*=8	*K*=6,8	*K*=8	*K*=6,8
	5.6	5.6	5.6	11.1	5.6	11.1
Totales	11.2		16.7		16.7	
Lognormal			*K*=8	*K*=6,8	*K*=8	*K*=6,8
			5.6	11.1	5.6	11.1
Totales			16.7		16.7	

Se estimaron los valores *p* asociados a los efectos fijos, con la aproximación KR, y las tasas de error Tipo I empíricas para cada combinación de las distintas variables del estudio.

La Tabla 4 muestra las tasas de error Tipo I empíricas para el efecto tiempo, así como los porcentajes de robustez. Con *K* = 4 y ε = 0.57, el procedimiento KR es robusto en un 66.7% cuando la distribución es normal, En ningún caso la prueba es robusta cuando la distribución es lognormal. Con *K* = 4 y ε = 0.75 la diferencia entre distribuciones normales y lognormales no es tan acusada.

Tabla 4. Tasas de error Tipo I empíricas del efecto tiempo (valor nominal de 0.05)

N	N_1	n_2	n_3	Δn_j	Cov. entre grupo	Empare-jamiento	Distribuciones			
							Normal		Lognormal	
							Esfericidad			
							0.57	0.75	0.57	0.75
					$K = 4$					
30	10	10	10	0.00	=	Nulo	0.075	0.061	**0.086**	**0.084**
36	12	12	12	0.00	=	Nulo	0.061	0.061	**0.085**	**0.087**
42	14	14	14	0.00	=	Nulo	0.066	0.047	**0.083**	**0.092**
30	10	10	10	0.00	≠	Nulo	0.067	**0.079**	**0.107**	0.068
36	12	12	12	0.00	≠	Nulo	0.064	0.064	**0.084**	0.068
42	14	14	14	0.00	≠	Nulo	0.068	0.067	**0.104**	0.064
30	5	10	15	0.41	≠	+	**0.079**	0.074	**0.091**	**0.078**
36	6	12	18	0.41	≠	+	0.072	0.073	**0.112**	0.075
42	7	14	21	0.41	≠	+	0.063	0.068	**0.107**	0.065
30	15	10	5	0.41	≠	–	**0.127**	**0.130**	**0.135**	**0.114**
36	18	12	6	0.41	≠	–	**0.092**	**0.096**	**0.126**	**0.090**
42	21	14	7	0.41	≠	–	**0.094**	**0.087**	**0.155**	**0.080**
					Porcentajes de robustez		66.7	66.7	0.0	41.7
					$K = 6$					
30	10	10	10	0.00	=	Nulo	0.075	0.070	**0.100**	**0.103**
36	12	12	12	0.00	=	Nulo	**0.079**	0.058	**0.111**	0.075
42	14	14	14	0.00	=	Nulo	0.064	0.073	**0.090**	**0.102**
30	10	10	10	0.00	≠	Nulo	0.054	0.065	**0.103**	**0.104**
36	12	12	12	0.00	≠	Nulo	0.054	0.070	**0.109**	**0.102**
42	14	14	14	0.00	≠	Nulo	0.068	0.060	**0.090**	**0.107**
30	5	10	15	0.41	≠	+	*0.017*	**0.079**	0.038	**0.086**
36	6	12	18	0.41	≠	+	*0.011*	0.062	*0.021*	0.067
42	7	14	21	0.41	≠	+	*0.011*	0.054	*0.017*	0.069
30	15	10	5	0.41	≠	–	0.319	0.083	0.345	0.113
36	18	12	6	0.41	≠	–	**0.277**	0.069	**0.301**	**0.125**
42	21	14	7	0.41	≠	–	0.272	0.088	0.320	0.076
					Porcentajes de robustez		41.7	75.0	8.3	25.0
					$K = 8$					
30	10	10	10	0.00	=	Nulo	**0.094**	0.072	**0.129**	**0.108**
36	12	12	12	0.00	=	Nulo	0.073	0.059	**0.111**	**0.102**
42	14	14	14	0.00	=	Nulo	0.073	0.057	0.115	0.098
30	10	10	10	0.00	≠	Nulo	0.073	**0.078**	**0.132**	**0.139**
36	12	12	12	0.00	≠	Nulo	0.073	0.055	**0.115**	**0.127**
42	14	14	14	0.00	≠	Nulo	0.069	0.075	**0.132**	**0.098**
30	5	10	15	0.41	≠	+	0.085	0.081	0.081	0.090
36	6	12	18	0.41	≠	+	0.066	0.071	0.099	0.090
42	7	14	21	0.41	≠	+	**0.087**	0.071	**0.080**	**0.079**
30	15	10	5	0.41	≠	–	0.094	0.103	0.123	0.131
36	18	12	6	0.41	≠	–	**0.092**	0.072	**0.124**	**0.119**
42	21	14	7	0.41	≠	–	**0.076**	0.075	**0.097**	**0.101**
					Porcentajes de robustez		50.0	75.0	0.0	0.0

Nota: En negrita = liberal; en cursiva = conservador.

Con $K = 6$ y $\varepsilon = 0.57$, el porcentaje de robustez es del 41.7% con distribucisn normal y con distribucisn lognormal disminuye a un 8.3%. Con $\varepsilon = 0.75$, el porcentaje de robustez es del 75% con distribución normal y del 25% con distribución lognormal, al ser la prueba liberal con emparejamientos nulos. Obsérvese que con emparejamientos positivos y $\varepsilon = 0.57$, la prueba es conservadora, tanto con distribuciones normales como lognormales.

Con $K = 8$ y $\varepsilon = 0.57$, la robustez es del 50% con distribución normal. En cambio, con distribución lognormal es del 0%. Donde se observa todavía una mayor diferencia es con $\varepsilon = 0.75$.

En la Tabla 5 se muestran las tasas de error Tipo I empíricas y los porcentajes de robustez del efecto de interacción tiempo x grupo. Con $K = 4$, se observa que apenas hay diferencias entre distribuciones normales y lognormales. Cuando $\varepsilon = 0.57$ el porcentaje es del 58.3% con distribución normal y del 50% con distribución lognormal. De forma similar, los porcentajes se mantienen cuando $\varepsilon = 0.75$. Así, el porcentaje de robustez es del 41.7% para distribución normal y lognormal. Del mismo modo, con $K = 6$ y $\varepsilon = 0.75$ no hay diferencias entre distribución normal y lognormal. Cuando la distribución es lognormal y $\varepsilon = 0.57$, el porcentaje de robustez es del 33.3%, frente al 25% cuando la distribución es normal.

Con $K = 8$, la prueba es más robusta cuando la distribución es lognormal. La mayor diferencia entre distribución normal y lognormal es con emparejamientos positivos, donde se observa que la prueba es robusta con distribución lognormal y no con distribución normal.

Tabla 5. Tasas de error Tipo I empíricas del efecto interacción (valor nominal de 0.05)

N	N_1	n_2	n_3	Δn_j	Cov. entre grupo	Empare-jamiento	Distribuciones			
							Normal		Lognormal	
							Esfericidad			
							0.57	0.75	0.57	0.75
					$K = 4$					
30	10	10	10	0.00	=	Nulo	**0.077**	**0.083**	0.063	0.065
36	12	12	12	0.00	=	Nulo	0.066	0.065	0.069	**0.076**
42	14	14	14	0.00	=	Nulo	0.073	0.066	0.071	0.075
30	10	10	10	0.00	≠	Nulo	0.073	**0.086**	**0.088**	**0.089**
36	12	12	12	0.00	≠	Nulo	0.063	**0.080**	0.075	0.075
42	14	14	14	0.00	≠	Nulo	0.064	0.069	**0.089**	0.072
30	5	10	15	0.41	≠	+	**0.088**	**0.094**	0.060	**0.093**
36	6	12	18	0.41	≠	+	0.072	0.069	**0.079**	0.074
42	7	14	21	0.41	≠	+	0.058	0.064	0.055	**0.081**
30	15	10	5	0.41	≠	–	0.171	0.198	0.189	0.187
36	18	12	6	0.41	≠	–	0.118	0.129	0.181	0.132
42	21	14	7	0.41	≠	–	0.111	0.111	0.191	0.121
					Porcentajes de robustez		58.3	41.7	50.0	41.7
					$K = 6$					
30	10	10	10	0.00	=	Nulo	0.094	0.083	0.081	0.083
36	12	12	12	0.00	=	Nulo	**0.082**	0.066	0.069	0.075
42	14	14	14	0.00	=	Nulo	0.063	0.061	**0.080**	0.064
30	10	10	10	0.00	≠	Nulo	**0.093**	0.059	0.111	0.108
36	12	12	12	0.00	≠	Nulo	0.083	0.083	0.100	0.097
42	14	14	14	0.00	≠	Nulo	**0.084**	0.065	**0.076**	**0.107**
30	5	10	15	0.41	≠	+	0.028	**0.092**	0.029	0.062
36	6	12	18	0.41	≠	+	0.025	0.068	0.030	0.062
42	7	14	21	0.41	≠	+	*0.020*	0.059	0.031	0.060
30	15	10	5	0.41	≠	–	0.350	0.116	0.357	0.119
36	18	12	6	0.41	≠	–	0.306	0.096	0.306	0.114
42	21	14	7	0.41	≠	–	0.308	0.095	0.310	0.087
					Porcentajes de robustez		25.0	50.0	33.3	41.7
					$K = 8$					
30	10	10	10	0.00	=	Nulo	**0.093**	**0.076**	0.075	**0.089**
36	12	12	12	0.00	=	Nulo	0.085	0.076	**0.076**	0.075
42	14	14	14	0.00	=	Nulo	0.068	0.071	**0.078**	**0.080**
30	10	10	10	0.00	≠	Nulo	0.083	0.119	0.115	0.103
36	12	12	12	0.00	≠	Nulo	0.103	0.082	0.095	0.123
42	14	14	14	0.00	≠	Nulo	0.086	0.080	0.096	0.105
30	5	10	15	0.41	≠	+	**0.094**	0.075	0.068	0.072
36	6	12	18	0.41	≠	+	**0.085**	**0.093**	0.068	0.051
42	7	14	21	0.41	≠	+	**0.086**	**0.083**	0.063	0.061
30	15	10	5	0.41	≠	–	0.104	0.116	0.128	0.130
36	18	12	6	0.41	≠	–	0.105	0.112	0.126	0.106
42	21	14	7	0.41	≠	–	0.082	0.105	0.108	0.096
					Porcentajes de robustez		8.3	16.7	33.3	33.3

Nota: En negrita = liberal; en cursiva = conservador.

DISCUSIÓN

El propósito básico de este estudio fue obtener la robustez del LMM en diseños longitudinales mixtos cuando los datos no se distribuyen normalmente. Un resultado interesante, al seleccionar la estructura de covarianza con matrices de covarianza entre-grupos heterogéneas, es el mejor ajuste de la matriz ARH_j, independientemente del tipo de distribución. Ello se observa con emparejamientos positivos y negativos. Esta estructura ARH_j de dependencia entre los datos de medidas repetidas, ya fue hallada con anterioridad por Keselman, Algina, Kowalchuk y Wolfinger (1998).

Por lo que respecta a las tasas de error Tipo I, puede afirmarse que para el efecto tiempo, la prueba es más robusta con distribuciones normales. La interacción da un incremento considerable del porcentaje de robustez de la prueba cuando la distribución es lognormal.

NOTA DE LOS AUTORES

Este estudio ha sido financiado por el Proyecto de Investigación PSI2009-11136 del Ministerio de Ciencia e Innovación.

REFERENCIAS

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532. doi:10.1007/BF02293811

Keselman, H. J., Algina, J., Kowalchuk R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics: Simulation and Computation, 27,* 591-604. doi:10.1080/03610919808813497

SAS Institute Inc. (2008). *SAS/STAT 9.2 user's guide*. Cary, NC: Author.

MODELO LINEAL MIXTO EN DISEÑOS *SPLIT-PLOT* CON CORRECCIÓN KENWARD-ROGER DE LOS GRADOS DE LIBERTAD. ESTUDIO DE SIMULACIÓN CON DATOS NORMALES Y NO NORMALES EN LOS GRUPOS

María J. Blanca y Rebecca Bendayan
Universidad de Málaga

Correo electrónico: blamen@uma.es

Resumen

El objetivo del presente trabajo es evaluar la robustez del modelo lineal mixto, con el ajuste de los grados de libertad mediante el método de Kenward-Roger, para analizar diseños *split-plot* con tamaños muestrales pequeños y cuando el supuesto de normalidad se satisface en todos los grupos o no se satisface con diferente grado de violación en los mismos. Se realiza un estudio de simulación Monte Carlo considerando un diseño *split-plot* con 3 grupos y 4 ocasiones de medidas repetidas, asumiendo una matriz de covarianza no estructurada en la generación de datos, bajo las siguientes condiciones: a) con grupos balanceados y no balanceados; b) homogeneidad y heterogeneidad de la matriz de covarianza; c) emparejamiento nulo, positivo o negativo entre el tamaño de grupo y la matriz de covarianza; d) esfericidad de 0,57 ó 0,75 y e) distribución normal o no normal de la variable dependiente. Se analizan y comparan las tasas de error Tipo I hallándose que, en general, hay una tendencia a la liberalidad en las condiciones estudiadas. Asimismo, el efecto de la no normalidad en el procedimiento K-R parece estar estrechamente ligado con la existencia de homogeneidad o no de varianzas.

El diseño *split-plot* o mixto es el más seguido en las publicaciones científicas españolas de psicología de la última década (Fernández, Vallejo, Livacic-Rojas y Tuero, 2010). El análisis de varianza mixto es la técnica analítica asociada a este tipo de diseño, partiendo de los supuestos de normalidad, independencia entre las observaciones y esfericidad multimuestra. Cuando las asunciones del modelo univariado se satisfacen, el ANOVA proporciona una prueba adecuada (Huynh y Feldt, 1970; Rouanet y Lepine, 1970), pero cuando no se satisfacen la robustez del procedimiento se ve comprometida (Balluerka y Vergara, 2004).

Se han desarrollado numerosos estudios con el fin de proporcionar pruebas estadísticas alternativas que sean robustas a las violaciones de los citados supuestos, ya

que éstas son muy frecuentes en la realidad investigadora (Keselman, Algina, Kowalchuk y Wolfinger, 1998; Fernández et al., 2010; Lei y Lomax, 2005; Micceri, 1989). Entre ellas es destacable el modelo lineal mixto (MLM; Laird y Ware (1982). Este modelo utiliza estadísticos de tipo Wald que son válidos con tamaños muestrales grandes pero con muestras pequeñas el error Tipo I suele presentar un sesgo positivo (Wright y Wolfinger, 1996). Para estos casos se ha propuesto el procedimiento de ajuste de los grados de libertad desarrollado por Kenward y Roger (K-R; 1997).

Los estudios de simulación muestran que el procedimiento K-R es robusto para el efecto principal de medidas repetidas, con esfericidad asumida, ante violaciones de la homogeneidad de varianzas y distintas violaciones de la normalidad (Kowalchuk, Keselman, Algina y Wolfinger, 2004; Livacic-Rojas, Vallejo y Fernández, 2006, 2010; Vallejo, Fernández, Herrero y Conejo, 2004). En cuanto al efecto de interacción, con esfericidad asumida, se ha encontrado que K-R es robusto (Kowalchuk et al., 2004; Livacic-Rojas, Vallejo y Fernández., 2010), conservador (Livacic-Rojas et al., 2006; Vallejo et al., 2004) o liberal (Vallejo y Ato, 2006). Escasean los estudios que analizan el efecto de la violación conjunta de la esfericidad y normalidad, encontrándose una tendencia a la liberalidad asociada a determinadas condiciones de heterogeneidad (Vallejo y Ato, 2006). En este tipo de estudios se suele suponer que todos los grupos presentan la misma distribución, aunque en investigaciones reales es probable que esto no suceda. De hecho, esta condición ha sido estudiada en el ANOVA intersujeto, encontrándose que el error Tipo I incrementa cuando los grupos tienen distintas distribuciones no normales (Glass, Peckham y Sanders, 1972; Harwell, Rubinstein, Hayes y Old, 1992; Lix, Keselman y Keselman, 1996).

El objetivo de este estudio es evaluar la robustez, mediante la tasa de error Tipo I, del MLM con el procedimiento K-R en diseños *split-plot* de muestras pequeñas ante violaciones de la normalidad en diferente grado en los distintos grupos, violaciones de la esfericidad y de la homogeneidad de varianza.

MÉTODO

Se ha llevado a cabo un estudio de simulación Monte Carlo considerando un diseño *split-plot* con un factor intersujeto con tres niveles (J=3) y un factor intrasujeto con cuatro ocasiones de medida (K=4). Para generar los datos se empleó un patrón de matriz de covarianza desestructurado (UN) y se utilizó el procedimiento de Fleishman (1978) y generalizado al caso multivariado por Vale y Maurelli (1983), utilizando el sistema SAS/IML. Las variables de estudio fueron:

Tamaño muestral de los grupos. El tamaño total de la muestra fue de 30, manipulando el tamaño de grupo, con diseños balanceados y no balanceados, siendo el coeficiente de variación, Δn_j, de 0 con tamaños de grupos iguales y de Δn_j = 0,41 con tamaños diferentes.

Forma de la distribución de la variable medida. Además de la distribución normal se incluyeron distribuciones con índices de asimetría y curtosis que no se

ajustan a ninguna distribución conocida, generándose una distinta para cada grupo. El valor de la asimetría se estableció en 0,8, y los valores de curtosis fueron para cada grupo 0,4, 2,4 y 5,4, respectivamente.

Esfericidad. Se consideraron valores de ε de 0,57 y 0,75, considerándose éste último como aproximación a la esfericidad (Arnau, Bono y Vallejo, 2009; Keselman, Algina, Kowalchuk y Wolfinger, 1999).

Igualdad de las matrices de dispersión. Se incluyeron matrices de covarianza grupales homogéneas y heterogéneas (1:3:5).

Emparejamiento de las matrices de covarianza y el tamaño de los grupos. En diseños balanceados, el emparejamiento fue nulo. En diseños no balanceados, el emparejamiento era positivo si el grupo de menor tamaño se asociaba con la matriz de dispersión menor, o negativo, en caso contrario.

Se analizan las tasas de error Tipo I asociadas a los efectos de medidas repetidas e interacción del MLM de cada una de las condiciones del estudio. Para cada combinación se han realizado 10.000 réplicas.

RESULTADOS

Para evaluar la robustez del MLM se utilizó el criterio de robustez de Bradley (1978) que considera una prueba robusta si la tasa empírica de error Tipo I se encuentra comprendida dentro del intervalo 0,025 y 0,075. Se considera liberal cuando ésta es mayor que el límite superior y conservadora, cuando se sitúa por debajo del límite inferior.

Distribución normal en todos los grupos. Como se observa en la Tabla 1, para el efecto de medidas repetidas el procedimiento K-R es robusto ante violaciones de la esfericidad, siempre que el tamaño de los grupos sea igual. Cuando existe un emparejamiento positivo entre la matriz de covarianza y el tamaño del grupo es liberal cuando se viola el supuesto de esfericidad; sin embargo, cuando el emparejamiento es negativo es liberal se cumpla o no el supuesto de esfericidad. Para el efecto de interacción el procedimiento se muestra liberal en todas las condiciones.

Distribución no normal y con diferente grado en todos los grupos. Los resultados (Tabla 2) muestran que al igual que ocurre cuando los datos son normales, para el efecto de medidas repetidas el procedimiento K-R es robusto ante violaciones de la esfericidad sólo cuando el tamaño de los grupos es igual y hay heterogeneidad de varianza. Cuando existe un emparejamiento positivo entre la matriz de covarianza y el tamaño del grupo este procedimiento es liberal cuando se viola el supuesto de esfericidad; sin embargo, cuando el emparejamiento es negativo es liberal se cumpla o no el supuesto de esfericidad. Es destacable que la violación del supuesto de normalidad afecta a este procedimiento, haciéndolo liberal cuando hay homogeneidad de varianzas. Para el efecto de interacción el procedimiento se muestra liberal en casi todas las condiciones, excepto cuando el emparejamiento es nulo y hay heterogeneidad de varianzas dónde el procedimiento se muestra robusto.

Tabla 1. Tasas de error Tipo I estimadas para los efectos tiempo e interacción en condiciones de normalidad (valor nominal 0,05)

N	n_1	n_2	n_3	Δn_j	**Covarianza entre grupos**	**Empare-jamiento**	**g1: γ_1= 0 γ_2= 0 g2: γ_1= 0 γ_2= 0 g3: γ_1= 0 γ_2= 0**	
							Esfericidad	
							0.57	**0.75**
Efecto de medidas repetidas								
30	10	10	10	0.00	=	Nulo	0.074	0.070
30	10	10	10	0.00	≠	Nulo	0.065	0.070
30	5	10	15	0.41	≠	+	**0.078**	0.075
30	15	10	5	0.41	≠	–	**0.120**	**0.125**
Efecto de interacción								
30	10	10	10	0.00	=	Nulo	**0.079**	**0.076**
30	10	10	10	0.00	≠	Nulo	**0.080**	**0.084**
30	5	10	15	0.41	≠	+	**0.091**	**0.087**
30	15	10	5	0.41	≠	–	0.176	0.185

Nota: N: tamaño muestral total; n_j : tamaño muestral de cada grupo. Δn_j : Coeficiente de variación. g_j : grupo. γ_1 : asimetría. γ_2 : curtosis. + - : emparejamiento curtosis-tamaño grupo positivo y negativo. En **negrita** : liberal.

Tabla 2. Tasas de error Tipo I estimadas para los efectos tiempo e interacción en condiciones de no normalidad (valor nominal 0,05)

N	n_1	n_2	n_3	Δn_j	**Covarianza entre grupos**	**Empare-jamiento**	**NO NORMAL g1: γ_1= 0.8 γ_2= 0.4 g2: γ_1= 0.8 γ_2= 2.4 g3: γ_1= 0.8 γ_2= 5.4**	
							Esfericidad	
							0.57	**0.75**
Efecto de medidas repetidas								
30	10	10	10	0.00	=	Nulo	**0.077**	0.072
30	10	10	10	0.00	≠	Nulo	0.071	0.067
30	5	10	15	0.41	≠	+	**0.083**	0.073
30	15	10	5	0.41	≠	–	**0.108**	**0.115**
Efecto de interacción								
30	10	10	10	0.00	=	Nulo	**0.078**	**0.076**
30	10	10	10	0.00	≠	Nulo	0.070	0.065
30	5	10	15	0.41	≠	+	**0.088**	**0.088**
30	15	10	5	0.41	≠	–	**0.154**	**0.158**

Nota: N: tamaño muestral total; n_j : tamaño muestral de cada grupo. Δn_j : Coeficiente de variación. g_j : grupo. γ_1 : asimetría. γ_2 : curtosis. + - : emparejamiento curtosis-tamaño grupo positivo y negativo. En **negrita** : liberal.

DISCUSIÓN

El objetivo del presente trabajo fue evaluar la robustez del MLM con la corrección de los grados de libertad por el método K-R con diseños *split-plot* de tamaños de muestra pequeños ante violaciones de la normalidad en diferente grado en los distintos grupos, violaciones de la esfericidad y de la homogeneidad de varianza.

Los resultados obtenidos muestran que cuando los grupos tienen distintas distribuciones no normales hay una tendencia a la liberalidad como ocurre en el ANOVA intersujeto (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996). Más concretamente, el efecto de la no normalidad en el procedimiento K-R parece estar estrechamente ligado con la existencia de homogeneidad o no de varianzas. Estos resultados son parcialmente coincidentes con los hallados por Vallejo y Ato (2006) que hallaron que K-R es liberal para el efecto de interacción.

Por otro lado, los efectos de la violación de la esfericidad se dan al margen de que el supuesto de normalidad se satisfaga o no, y podría estar más asociado al tipo de emparejamiento entre la matriz de covarianza y el tamaño del grupo. Estos resultados son difícilmente comparables a los obtenidos en otros estudios ya que las condiciones estudiadas difieren de las utilizadas en el presente trabajo, haciéndose patente la necesidad de futuros estudios.

En conclusión, los resultados obtenidos muestran una clara tendencia a la liberalidad del procedimiento K-R ante violaciones de la normalidad, esfericidad y homogeneidad de varianza cuando el tamaño muestral es reducido, poniéndose de relieve la necesidad de profundizar en esta cuestión con otras condiciones de estudio, como por ejemplo, otros tamaños muestrales y distintas distribuciones no normales conocidas y desconocidas.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado por el Proyecto de Investigación PSI2009-11136 del Plan Nacional I+D+I del Ministerio de Ciencia e Innovación.

REFERENCIAS

Arnau, J., Bono, R. y Vallejo, G. (2009) Analyzing small samples of repeated measures data with the mixed-model adjusted F test. *Communications in Statistics. Simulation and Computations, 38*, 1083-1103.

Balluerka, N. y Vergara, A.I. (2004). *Diseño de investigación experimental en Psicología.* Madrid: Prentice Hall.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Fernández, P., Vallejo, G., Livacic-Rojas, P. y Tuero, E. (2010). Características y análisis de los diseños de medidas repetidas en la investigación experimental en España en los últimos 10 años. *Actas del XI Congreso de Metodología de las Ciencias del Comportamiento.* Málaga.

Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika, 43,* 4, 521-531.

Glass, G.V., Peckham, P.D. y Sanders, J.R. (1972). Consequences of failure to meets assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42,* 237-288.

Harwell, M.R., Rubinstein, E.N., Hayes, W.S. y Olds, C.C. (1992). Summarizing Monte Carlo results in methodological research: the one and two factors fixed effect ANOVA case. *Journal of Educational Statistics, 17,* 315-339.7

Huynh, H. y Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurement designs have exact F-Distribution. *Journal of the American Statistical Association, 65*, 1582-1589.

Kenward, M.G. y Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983–997.

Keselman, H.J., Algina, J., Kowalchuk, R.K. y Wolfinger, R.D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics – Computation and Simulation, 27*, 591-604.

Keselman, H.J., Algina, J., Kowalchuk, R.K. y Wolfinger, R.D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 52*, 63-78.

Kowalchuk, R.K., Keselman, H.J., Algina, J. y Wolfinger, R.D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement, 64,* 2, 224-242.

Laird, N.M. y Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics, 38,* 963-974.

Lei, M. y Lomax, R. G. (2005). The effect of varying degrees on nonnormality in structural equation modeling. *Structural equation modeling, 12*, 1-27.

Livacic-Rojas, P., Vallejo, G. y Fernández, P. (2006). Procedimientos estadísticos alternativos para evaluar la robustez mediante diseños de medidas repetidas. *Revista Latinoamericana de Psicología, 38,* 3, 579-598.

Livacic-Rojas, P., Vallejo, G. y Fernández, P. (2010). Analysis of Type I error rate of univariable and multivariate procedures in repeated measures designs. *Communications in Statistics - Simulation and Computation, 39*, 624-640.

Lix, L., Keselman, J.C. y Keselman, H.J. (1996). Consequences of assumptions violations revisited: A quantitative review of alternatives to the one-way analysis of variance *F* test. *Review of Educational Research, 66*, 579-620.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Rouanet, H. y Lepine, D. (1970). Comparison between treatments in a repeated measures design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23*, 147-163.

Vale, C.D. y Maurelli, V.A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48,* 3, 465-471.

Vallejo, G. y Ato, M. (2006). Modified Brown-Forsythe procedure for testing interaction effects in split-plot designs. *Multivariate Behavioral Research, 41,* 4, 549-578.

Vallejo, G. Fernández, P., Herrero F.J. y Conejo, N.M. (2004). Alternative procedures for testing fixed effects in repeated measures designs when assumptions are violated. *Psicothema, 16,* 3, 498-508.

Wright, S.P. y Wolfinger, R.D. (1996). Repeated measures analysis using mixed models: Some simulation results. *Conference on Modelling Longitudinal and Spatially Correlated Data:Methods, Applications, and Future Directions*. Nantucket, MA.

SELECCIÓN DE MODELOS MULTINIVEL USANDO EL CRITERIO DE INFORMACIÓN DE AKAIKE CONDICIONAL

G. Vallejo[1], M. P. Fernández[1], P. E. Livacic-Rojas[2] y E. Tuero-Herrero[1]
[1] Universidad de Oviedo
[2] Universidad de Santiago de Chile
Correo electrónico: gvallejo@uniovi.es

Resumen

El presente trabajo examina el desempeño de los criterios AIC's, marginal y condicional, en presencia de efectos aleatorios. Cuando se emplean criterios de selección para elegir entre modelos jerárquicos, uno no sólo tiene que decidir qué criterio usar, sino también especificar el procedimiento de estimación a emplear y la forma de computar los grados de libertad. En los modelos mixtos el interés se centra tanto en los efectos fijos, que se supone compartidos por toda la población, como en los efectos aleatorios, que se supone únicos para cada individuo. Por ejemplo, supongamos que se desee probar la eficacia de una nueva terapia para reducir los trastornos de ansiedad. Cuando la atención se centra en los efectos fijos se quiere saber si la tasa de cambio promedio difiere entre los grupos, mientras que cuando la atención se centra en los efectos aleatorios se quiere saber cómo cambian los perfiles individuales a lo largo del tiempo dentro de cada grupo. Bajo la primera situación, los criterios de selección que ignoran los efectos aleatorios, tales como el AIC o el BIC marginales, son apropiados. No obstante, cuando el interés se centra en los individuos, los efectos fijos se estiman condicionales a la presencia de componentes aleatorios y los criterios de selección que ignoran dichos componentes resultan inapropiados.

De acuerdo con Bono, Arnau y Vallejo (2008), en las ciencias sociales y del comportamiento los tradicionales métodos ANOVA y MANOVA siguen siendo dominantes para analizar diseños de medidas repetidas. Sin embargo, a pesar de su simplicidad técnica e interpretativa, estos métodos no siempre constituyen la mejor opción para analizar los datos obtenidos en el mundo real. En la investigación aplicada, además de la frecuente violación de los supuestos de normalidad, homogeneidad y esfericidad, resulta normal toparse con unidades experimentales de diferente tamaño, con datos desequilibrados e incompletos, con variables explicativas potencialmente cambiantes, con diferentes tipos de dependencia entre las medidas repetidas, con mediciones registradas en periodos temporales desigualmente espaciados y con diferentes tipos de cambio. La solución natural para intentar abordar los problemas planteados, la proporciona el modelo lineal mixto

(MLM) o, alternativamente, el MLM generalizado si la naturaleza de la variable de respuesta origina datos categóricos.

Con datos de corte longitudinal el MLM resultan muy útil, ya que permite modelar la estructura de medias (EM) y la estructura de covarianza (EC). Cuando se emplea el MLM, más que asumir EC excesivamente simples o completamente generales, se trata de buscar un equilibrio entre los criterios de parquedad y flexibilidad. Como han puesto de relieve Vallejo, Ato y Valdés (2008), especificando un modelo excesivamente simple se corre el riesgo de efectuar inferencias erróneas, debido a la subestimación de los errores estándar (ES), mientras que especificando un modelo excesivamente complejo se corre el riesgo de efectuar inferencias ineficientes. De hecho, en el trabajo de Vallejo et al. (2008), las tasas de error se mantuvieron próximas al valor nominal elegido cuando el verdadero proceso generador de los datos (VPGD) se especificaba correctamente; sin embargo, los ES resultaban sesgados cuando el VPGD se especificaba erróneamente. Aunque la selección del VPGD resulte central para interpretar correctamente los datos, dicho objetivo no es fácil de lograr, pues para una misma evidencia muestral existen múltiples modelos candidatos (Claeskens & Hjort, 2008).

Para facilitar la elección del mejor MLM, diversas técnicas de inferencia y herramientas de selección están disponibles. Para discriminar entre modelos anidados, resulta frecuente usar la prueba de razón de verosimilitudes (LRT) con la desvianza obtenida a partir de la función de máxima verosimilitud completa (FML) o restringida/residual (REML), según se trate de elegir entre modelos con idéntica EC o EM. Sin embargo, en el contexto de datos longitudinales con errores correlacionados, es común comparar modelos no anidados, en especial cuando se trata de elegir entre modelos con idéntica EM (Gurka & Edwards, 2008). En este caso, se recomienda realizar el proceso de selección usando criterios de información (CI; Azari, Li & Tsai, 2006; Shang & Cavanaugh, 2008; Lee & Ghost, 2009; Kitagawa & Konishi, 2010). Otros criterios de selección, tales como el coeficiente de determinación, el coeficiente de correlación de concordancia (CCC) o la suma de cuadrados residual de predicción (PRESS), han recibido escasa atención. No obstante lo anterior, Wang y Schaalje (2009) informan que el desempeño de los criterios predictivos CCC y PRESS no supera al desempeño de los IC.

En el contexto de los modelos anidados, Vallejo, Arnau, Bono, Fernández y Tuero (2010) encontraron que el desempeño del LRT basado en el estimador FML era superior al de los CI examinados. Sin embargo, el desempeño del LRT era inferior cuando se basaba en el estimador REML. Vallejo et al. (2010), también encontraron que los CI basados en el método REML elegían la verdadera EM tan bien o mejor que los CI basados en el método FML. Este resultado confirma los hallazgos encontrados por Gurka (2006) con modelos no anidados, en contra de lo defendido en la literatura estadística especializada.

A su vez, en el contexto de los modelos no anidados, Vallejo, Fernández, Livacic-Rojas y Tuero (2011), han examinado el desempeño de múltiples CI marginales a la hora seleccionar de forma conjunta la EM y la EC en presencia de datos incom-

pletos motivados por abandonos. Asimismo, el estudio investigó el impacto ocasionado por manipular diferentes formas de la distribución, métodos de estimación y ajustes en el cálculo de los CI. Aunque los beneficios de la aplicación de los CI son generalmente reconocidos como fundamentales para efectuar buenas inferencias, Vallejo et al. (2011) encontraron que todos los CI marginales examinados resultaban de utilidad limitada cuando EC complejas se combinaban con muestras pequeñas y cuando los datos se obtenían a partir de distribuciones moderadamente sesgadas.

Al margen del efecto negativo que los tamaños de muestra reducidos y las EC complejas ejercían en el desempeño de los CI, también conviene resaltar que los CI marginales han sido desarrollados en el contexto de los modelos de regresión clásica y paralelamente extendidos al ajuste de los MLM, lo cual favorece la selección de modelos sin efectos aleatorios. Esta deficiencia ha sido puesta de relieve por distintos investigadores, incluyendo Vaida y Blanchard (2005), Liang, Wu y Zou (2008), Greven y Kneib (2009) y Srivastava y Kubokawa (2010). Los autores citados han desarrollado diferentes versiones del criterio AIC derivándolas desde los estimadores FML y REML condicionales, en lugar de hacerlo desde los estimadores FML y REML marginales. Por consiguiente, en la presente comunicación se pretende examinar el desempeño del criterio AIC marginal (AICm) y AIC condicional (AICc) por desarrollado Vaida y Blanchard (2005) y esto lo haremos de una manera bastante restringida debido a la cantidad de programación exigida. Entre los múltiples programas estadísticos comerciales, sólo el módulo PROC GLIMMIXED del SAS (versión 9.2 TSM3, 2010) proporciona la función de FML condicional, no así la función de REML condicional ni los grados de libertad correspondientes a la función de penalización.

MÉTODO DE LA SIMULACIÓN

Para dar respuesta de los objetivos planteados llevamos acabo un estudio de simulación en que utilizamos un diseño cross-over con dos tratamientos, dos secuencias y múltiples periodos. En concreto, los participantes del primer grupo recibieron la secuencia de tratamiento 2(A)2(B), 3(A)3(B), 4(A)4(B) y 5(A)5(B), mientras que los del segundo grupo recibieron la secuencia inversa para contrarrestar los posibles efectos residuales.

(a) *Métodos de estimación*. La evaluación fue realizada bajo estimación FML y REML. Concretamente, la evaluación del desempeño implicaba seleccionar de un conjunto de diez modelos candidatos el VPGD. En la Tabla 1 aparecen recogidos los modelos utilizados en la comparación, así como el valor de los parámetros de efectos fijos usados para generar los datos.

(b) *Tamaño de muestra*. Los conglomerados usados tenían los tamaños siguientes: N = 20 (10-10), N = 40 (20-20), N = 60 (30-30), N = 80 (40-40) y N = 100 (50-50).

(c) *Número de medida repetidas*: Los periodos manipulados fueron: t = 4, 6, 8 y 10.

Tabla 1. Conjunto de modelos de candidatos y valor de los parámetros de efectos fijos

M_1	$E(y_{ij}) = \beta_{00}$
M_2	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j$
M_3	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet}$
M_4	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet}$
M_5	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet}$
M_6	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet}$
M_7	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet}$
$M_8^{©}$	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + \beta_{21}G_j \times CT_{ij}^{\bullet}$
M_9	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + \beta_{21}G_j \times CT_{ij}^{\bullet} + \beta_{30}T_{ij}^{2\bullet}$
M_{10}	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + \beta_{21}G_j \times CT_{ij}^{\bullet} + \beta_{20}T_{ij}^{2\bullet} + \beta_{31}G_j \times T_{ij}^{2\bullet}$
$\boldsymbol{\beta}' = [\beta_{00} = 1.00;\ \beta_{01} = 1.25;\ \beta_{10} = -0.50;\ \beta_{11} = 0.50;\ \beta_{20} = -0.50;\ \beta_{21} = 0.50]$	

Nota: $M_8^{©}$= modelo usado para generar los datos.

El desempeño de las herramientas de selección fue evaluado asumiendo que las matrices de covarianza grupales eran homogéneas y los datos se distribuían normalmente. Para realizar los cálculos se utilizó un MACRO escrito en lenguaje SAS/IML (2010) y las condiciones del diseño fueron replicadas 5000 veces usando el algoritmo Simplex en el proceso de optimización de la función de verosimilitud residual.

RESULTADOS

En la Tabla 2 que se muestra más abajo aparece tabulado el porcentaje de veces que los criterios examinados elegían correctamente las EC y EM, tanto bajo estimación FML como bajo estimación REML. Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables manipuladas en la investigación. Globalmente, los resultados indican lo siguiente:

- La ejecución de los criterios examinados dependía del método de estimación, tamaño de los conglomerados y número de medidas repetidas. Aunque no aparece recogido, el valor del vector de parámetros correspondientes a los efectos aleatorios lo hacía moderadamente.
- El desempeño de los CI fue ligeramente superior bajo estimación REML que bajo estimación FML. Promediando a través de las condiciones manipuladas, el porcentaje de aciertos obtenidos vía REML fue del 91.1%, mientras que el obtenido vía FML promediando a través de condiciones manipuladas fue del 90.3%.

– El AICc, tanto asintótico como para muestras finitas, tuvo un desempeño sustancialmente superior al del AICm bajo estimación FML, mientras que bajo estimación REML ocurrió lo contrario.

– Bajo estimación REML las diferencias promedio entre los AICm y AICc no excedían los 3 puntos porcentuales, mientras que bajo estimación FML excedían a 10. De hecho, bajo estimación REML, a partir de seis mediadas repetidas el desempeño del AICm y del AICc tendía a converger.

– Como se puede apreciar en la Tabla 2, a partir de 100 conglomerados el desempeño del AIC asintótico y el AIC corregido para muestras finitas, tanto marginal como condicional, era virtualmente idéntico.

Tabla 2. Porcentaje de veces que los AICm y AICc elegían correctamente el MLM bajo estimación FML y REML

N	FML				REML			
	AICCm	**AICm**	**AICCc**	**AICc**	**AICCm**	**AICm**	**AICCc**	**AICc**
MR: 4								
20	51.8	38.0	82.4	63.6	72.2	67.6	60.0	59.8
40	78.2	73.0	94.8	86.8	87.0	86.6	76.4	76.0
60	87.8	86.2	97.2	92.6	91.4	91.4	86.6	86.4
80	93.2	91.6	99.4	97.2	94.4	94.4	91.2	91.2
100	97.4	97.2	99.6	98.8	97.2	97.2	95.4	95.4
	81.7	**77.2**	**94.7**	**87.8**	**88.4**	**87.4**	**81.9**	**81.8**
MR: 6								
20	61.2	53.4	92.8	74.6	78.8	78.8	65.4	64.2
40	81.2	80.6	98.6	91.6	90.4	90.4	82.6	81.4
60	93.6	93.4	99.6	97.2	94.4	94.4	90.0	89.4
80	97.4	97.2	99.8	99.4	98.8	98.8	96.6	96.6
100	99.2	99.2	100.0	99.8	99.4	99.4	97.8	97.8
	86.5	**84.8**	**98.2**	**92.5**	**92.4**	**92.4**	**86.5**	**85.9**
MR: 8								
20	63.2	57.2	96.0	89.0	84.4	84.4	82.8	82.8
40	85.2	84.3	98.2	94.2	91.0	91.0	90.2	90.2
60	95.6	95.4	100.0	98.0	94.6	94.6	94.6	94.6
80	97.4	97.2	100.0	99.8	98.8	98.8	98.8	98.8
100	99.2	99.4	100.0	100.0	99.8	99.8	99.8	96.8
	88.1	**87.1**	**98.8**	**96.2**	**93.7**	**93.7**	**93.2**	**93.2**
MR: 10								
20	64.0	57.2	96.0	89.6	88.6	88.6	88.4	88.4
40	85.8	84.6	98.8	94.6	95.2	95.2	95.0	95.0
60	95.8	96.0	100.0	98.0	98.2	98.2	98.2	98.2
80	98.0	98.2	100.0	99.8	99.0	99.0	99.0	99.0
100	99.8	99.4	100.0	100.0	99.6	99.6	99.6	99.6
	88.7	**87.1**	**99.0**	**96.4**	**96.1**	**96.1**	**96.0**	**96.0**

CONCLUSIONES Y RECOMENDACIONES

Globalmente, el AICc trabajaba mejor que el AICm cuando el método utilizado en la estimación fue FML (95.5% *versus* 85.1%) y, viceversa, cuando se usó el método REML (92.5% *versus* 89.5%). En cualquier caso, con independencia del método de estimación empleado, el desempeño del AICc resultó ser ligeramente superior al del AICm. No obstante, conviene señalar que AICm lo proporcionan por defecto la mayor parte de los programas comerciales existentes, mientras que la obtención del AICc resulta excesivamente pesado. Hecha esta aclaración, debemos advertir que los resultados son limitados a las condiciones examinadas, si bien conjeturamos que pueden ser generalizadas a un rango más amplio de condiciones; por ejemplo, a situaciones donde los modelos no se hallen anidados unos dentro de otros, a situaciones donde la varianza de las observaciones sea heterogénea y la correlación entre las mismas decrezca a lo largo del tiempo y a situaciones donde las puntuaciones no se distribuyan normalmente. Finalmente, en la investigación realizada el VPGD siempre pertenecía a la familia de modelos investigados. No obstante, cuando se trabaja con datos reales desconocemos si el VPGD pertenece a la clase de modelos considerados. Por lo tanto, sería deseable realizar una investigación donde el objetivo fuese comparar el desempeño del AICm y AICc en términos de seleccionar el modelo más próximo al VPGD, dado que éste no se haya incluido en el conjunto de modelos presentes en la comparación.

NOTA DE LOS AUTORES

Este trabajo ha sido financiado mediante el proyecto de investigación concedido por el MCI (Ref.: PSI-2008-03624).

REFERENCIAS

Azari, R., Li, L., & Tsai, C.L. (2006). Longitudinal data model selection. *Computational Statistics and Data Analysis*, *50,* 3053–3066.

Bono, R., Arnau, J. & Vallejo, G. (2008). Técnicas de análisis aplicadas a datos longitudinales en psicología y ciencias de la salud: Período 1985-2005. *Papeles del Psicólogo, 29,* 1-15.

Claeskens, G., & Hjort, N.L. (2008). *Model Selection and Model Averaging*. New York, NY: Cambridge University Press.

Greven, S., & Kneib T. (2009). On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models. Technical report, paper 179. Johns Hopkins University, School of Public Health.

Gurka, M.J. (2006). Selecting the best linear mixed model under REML. *The American Statistician, 60*, 19-26.

Gurka, M.J., & Edwards, L.J. (2008). Mixed models. En C.R. Rao, J.P. Miller & D.C. Rao (Eds.): *Handbook of Statistics, Vol 27, Epidemiological and Medical Statistics* (pp. 253-280). New York, NY: Elsevier.

Kitagawa, G., & Konishi, S. (2010). Bias and variance reduction techniques for bootstrap information criteria. *Annals of the Institute of Statistical Mathematics, 62*, 209–234.

Lee H., & Ghosh, S.K. (2009). Performance of information criteria for spatial models. *Journal of Statistical Computation and Simulation*, *79*, 93-106.

Liang, H., Wu, H., & Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika, 95*, 773- 778.

SAS Institute Inc. (2010). *SAS/STAT ® 9.2 TSM3 user's guide.* Cary, NC: Author.

Shang, J., & Cavanaugh, J.E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational Statistics and Data Analysis, 52,* 2004–2021

Srivastava, M., & Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis, 101*, 1970-1980.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*, 351-370.

Vallejo, G., Arnau, J., Bono, R., Fernández, P., & Tuero, E. (2010). Selección de modelos anidados para datos longitudinales usando criterios de información y la estrategia de ajuste condicional. *Psicothema, 22*, 323-333.

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology. 4,* 10-21.

Vallejo, G., Fernández, P., Livacic-Rojas, P., & Tuero, E. (2011). Selecting the best covariance pattern regression model with missing data. *Behavior Research Methods, 43*, 18-36.

Wang, J., & Schaalje, G.B. (2009). Model selection for linear mixed models using predictive criteria. *Communications in Statistics - Simulation and Computation, 38*, 788-801.

VALIDEZ: NUEVAS APROXIMACIONES METODOLÓGICAS PARA NUEVOS DESAFÍOS

Coordinadores: Juana Gómez-Benito[1] y José Luis Padilla[2]
[1] Universidad de Barcelona
[2] Universidad de Granada

Ya en las dos últimas décadas del siglo pasado, se extendió un amplio consenso entre investigadores y profesionales sobre el papel central de la validez en la evaluación de la calidad de las mediciones aportadas por tests y cuestionarios. La relación de fuentes de evidencias de validez incluidas en la última edición de los Standards (AERA, APA, NCME, 1999), los fundamentos conceptuales propuestos por Samuel Messick y la aproximación basada en argumentos de Michael Kane, configuran los estudios de validación como diseños de investigación pensados para aportar evidencias que apoyen o refuten los supuestos que subyacen a las interpretaciones de las mediciones. Dirigidos a la búsqueda de evidencias, los estudios de validación están abiertos a gran variedad de aproximaciones metodológicas: estudios experimentales, modelos de ecuaciones estructurales, juicios de expertos, grupos focales, entrevistas cognitivas, etc. El objetivo principal del simposio es compartir experiencias en el desarrollo de estudios de validación desde una pluralidad de aproximaciones metodológicas. El simposio pretende reunir investigadores y profesionales motivados por los nuevos desafíos que plantean los estudios de validación, incluyendo consideraciones sobre las consecuencias del uso de los tests y cuestionarios en los procesos de tomas de decisión.

PALABRAS CLAVE: Validez, Elaboración/adaptación de tests, Juicio de expertos, Entrevista cognitiva.

«PEOPLE PUSH ME TO PLAY». ADAPTACIÓN DE UN CUESTIONARIO DE MOTIVACIÓN EN PSICOLOGÍA DEL DEPORTE

Carme Viladrich, Jaume Cruz y Miquel Torregrosa
Universidad Autónoma de Barcelona
Correo electrónico: carme.viladrich@uab.cat

Resumen

En el proceso de adaptación del *Behavioral Regulation in Sports Questionnaire* al idioma español se han redactado y evaluado cinco versiones sucesivamente más refinadas del cuestionario, basadas en las opiniones de personas expertas en lingüística y traducción, en psicología del deporte, y en metodología, además de las opiniones y las respuestas cuantitativas de jóvenes deportistas pertenecientes a la población diana a la que se dirige el cuestionario. Se presentan y discuten los datos relativos a dos de los ítems que han resultado particularmente informativos sobre el proceso de adaptación. Los resultados muestran que la adaptación de un cuestionario mejora de forma clara cuando se triangulan datos procedentes de voces diversas y obtenidos con metodologías variadas.

El Cuestionario de Regulación Conductual en el Deporte (BRSQ, por sus siglas en inglés *Behavioral Regulation in Sport Questionnaire*; Lonsdale, Hodge y Rose, 2008) se diseñó con el fin de proporcionar medidas de los seis tipos de regulación conductual previstos en la teoría de la autodeterminación (SDT, por su denominación en inglés *Self Determination Theory*; Deci y Ryan, 1985). Según dicha teoría, pueden distinguirse tres tipos de motivación, la intrínseca, la extrínseca y la amotivación. A su vez, la motivación extrínseca se manifiesta en cuatro tipos de regulaciones ordenadas de menos a más autónomas, a saber, la regulación externa, la introyectada, la identificada y la integrada. Una persona muestra motivación intrínseca cuando realiza una actividad por sí misma; un ítem típico de este tipo de motivación es «practico este deporte porque disfruto». En el otro extremo, una persona muestra amotivación cuando no identifica motivos por los que realiza una actividad; un ítem típico sería «practico este deporte, a pesar de que me pregunto porque paso por esto». Entre ambos extremos, cuando la regulación es externa, las actividades se realizan por las gratificaciones provenientes del exterior, como por ejemplo «para satisfacer a las personas que quieren que lo practique»; cuando es introyectada, se realizan para evitar sensación de culpa o para realzar el ego (e.g., «porque me avergonzaría dejarlo»); si la regulación es identificada, por lo que

aportan a la persona (e.g., «porque me enseña autodisciplina»); y finalmente, si es integrada, como parte de la forma de ser de la persona (e.g., «porque es una oportunidad de ser quien realmente soy») . Se espera que la persona con puntuaciones muy elevadas en las formas de regulación más autónomas, tenga puntuaciones más bajas en las formas de regulación controladas y en amotivación.

Puesto que la regulación conductual es específica para cada dominio comportamental (Vallerand, 2007), el BRSQ cumple con la función de evaluarla en el dominio de la práctica del deporte mediante cuatro ítems por cada uno de los seis tipos de regulación. Tanto la versión original (Lonsdale et al. 2008) como la adaptación española (Viladrich, Torregrosa y Cruz, 2011) han mostrado poseer propiedades psicométricas adecuadas. El método y los resultados de esta adaptación se han descrito de forma global en el citado artículo de Viladrich et al.

De forma más concreta y detallada, en el presente trabajo pretendemos ilustrar los pormenores del proceso de adaptación del BRSQ al español, centrándonos en la evolución de los dos ítems del cuestionario que resultaron particularmente informativos. El primero de ellos es «I participate in my sport because people push me to play» y forma parte de la subescala de regulación externa; y el segundo es «I participate in my sport because I feel obligated to continue» y forma parte de la subescala de regulación integrada.

MÉTODO

Participantes

En la fase de adaptación lingüística, participaron seis psicólogos, dos traductoras profesionales, y 27 deportistas federados (30% chicas) de entre 12 y 18 años.

En el primer estudio cuantitativo, participaron 578 deportistas (M=14,04 años; D.T.=1,71; 27% chicas). Todos entrenaban y competían regularmente en deportes federados a nivel comarcal y/o nacional, individuales o de equipo. La muestra fue incidental, obtenida en 2009 y 2010.

En el segundo estudio cuantitativo participaron 169 deportistas (M=20,65 años; D.T.=2,93; 17% chicas). Todos practicaban deportes individuales o de equipo, competían a nivel regional, nacional o internacional, y eran estudiantes en el Grado de Ciencias de la Actividad Física y del Deporte en centros universitarios nacionales a finales del año 2010.

Instrumento

Cuestionario de Regulación Conductual en el Deporte (BRSQ, *Behavioral Regulation in Sport Questionnaire*; Lonsdale et al., 2008), diseñado para evaluar la motivación en la práctica del deporte desde la perspectiva de la SDT. Está formado por seis subescalas, de cuatro ítems cada una, para medir la amotivación, las regu-

laciones externa, introyectada, identificada e integrada, y la motivación intrínseca. Cada ítem se contesta en una escala tipo Likert, que va de 1 (*Completamente falso*) a 7 (*Completamente verdadero*). Todas las afirmaciones son directas, y la puntuación total de cada subescala se obtiene promediando las respuestas a sus cuatro ítems, de manera que una puntuación más elevada se interpreta como mayor regulación del tipo que mide la escala.

Procedimiento

Para la adaptación cultural y lingüística diseñamos una modificación del protocolo de Acquadro, Conway, Giroudet y Mear (2004), de acuerdo con las recomendaciones de la *International Test Comission* (Hambleton, 2005). La particularidad más destacable de nuestro protocolo son las sucesivas reuniones del comité de expertos para redactar versiones cada vez más refinadas del cuestionario, incorporando la información proporcionada por traductores profesionales, especialistas en psicología del deporte y en metodología, y personas pertenecientes a la población diana. Se redactaron cinco versiones del cuestionario hasta considerar que las distintas informaciones convergían suficientemente.

Para la obtención de datos cuantitativos, utilizamos el protocolo descrito en Ramis, Torregrosa, Viladrich y Cruz (2010) del que destacamos los aspectos de aceptación voluntaria a colaborar por parte de deportistas, entrenadores e instituciones, garantía de confidencialidad de los datos individuales, y atención por parte de dos investigadores a lo largo de toda la sesión.

RESULTADOS

De acuerdo con el objetivo del presente trabajo, se presentan las decisiones lingüísticas y los datos cuantitativos relativos a dos ítems en las versiones 4 y 5 del cuestionario.

En la Tabla 1 se presentan los datos derivados de las decisiones lingüísticas, en la versión del cuestionario para chicas. En la redacción de la versión 4, la discusión del comité de expertos se centró en el ítem «because people push me to play» de la escala de regulación externa. Para el término «people» las opciones fueron «alguna gente» y «los demás». Se eligió «los demás» por considerarlo una expresión más natural. El término «push me» había sido traducido como «me animan» y también como «me empujan». Se aceptó «me empujan» porque a pesar de tratarse de una expresión poco natural en castellano, evitaba incluir un sesgo positivo que no estaba presente en la redacción original del ítem. Los ítems de la escala de regulación introyectada se tradujeron sin dificultades. Sin embargo, al leerlos conjuntamente, el ítem «porque me siento obligada a continuar» se consideró ambiguo, puesto que si alguien interpretaba que el sentimiento de obligación era respecto a los demás, el ítem se convertía en un indicador de regulación externa en lugar serlo de regulación introyectada.

En la redacción de la versión 5, se propuso traducir la expresión «push me» por «me exigen». El comité consideró que el concepto de exigencia podía considerarse sin connotaciones negativas en el contexto deportivo. Por otra parte, el mismo concepto permitía también eliminar la ambigüedad de la expresión «me siento obligada a continuar» sustituyéndola por «me exijo continuar».

Tabla 1. Decisiones lingüísticas en la adaptación del BRSQ

Inglés	**Español** **Versión 4** *Cambios en la versión 5*
I participate in my sport…	Practico este deporte...
Externa	
because if I don't other people will not be pleased with me.	porque si no lo hago los demás estarán descontentos de mi.
because I feel pressure from other people to play.	porque me siento presionada por los demás para seguir haciéndolo.
because people push me to play.	porque los demás me empujan a hacerlo. *porque los demás me exigen hacerlo.*
in order to satisfy people who want me to play.	para satisfacer a las personas que quieren que lo practique.
Introyectada	
because I would feel ashamed if I quit.	porque me avergonzaría dejarlo.
because I would feel like a failure if I quit.	porque me sentiría fracasada si lo dejara.
because I feel obligated to continue.	porque me siento obligada a continuar. *porque me exijo continuar.*
because I would feel guilty if I quit.	porque me sentiría culpable si lo dejara.

Nota: El sombreado es para facilitar la lectura.

Tabla 2. Valores alfa de Cronbach y promedio de respuestas para cada una de las subescalas

Subescala	**Alfa**		**Promedio**	
	Versión 4	**Versión 5**	**Versión 4**	**Versión 5**
Amotivación	,710	,848	1,86	1,81
Externa	**,618**	**,822**	**1,76**	**1,58**
Introyectada	**,729**	**,665**	**2,05**	**2,71**
Identificada	,685	,741	5,17	5,14
Integrada	,746	,714	5,26	4,88
Intrínseca	,797	,777	6,44	6,32

Tabla 3. Estadísticos descriptivos y cargas factoriales de los ítems de las subescalas de regulación externa e introyectada del BRSQ

Contenido ítem	Media		D.T.		Carga factorial	
	V4	V5	V4	V5	V4	V5
Externa						
los demás descontentos	1,69	1,50	1,32	1,11	,682	,914
presionada por los demás	1,33	1,53	0,95	1,17	,827	,915
los demás me empujan	***1,75***		***1,41***		***,811***	
los demás me exigen		***1,34***		***0,92***		***,928***
satisfacer a los demás	2,26	1,95	1,77	1,54	,659	,612
Introyectada						
me avergonzaría dejarlo	1,94	1,86	1,72	1,55	,782	,651
me sentiría fracasada	2,04	2,11	1,76	1,68	,789	,758
me siento obligada	***1,89***		***1,62***		***,637***	
me exijo		***4,55***		***1,95***		***,521***
me sentiría culpable	2,33	2,32	1,93	1,71	,753	,767

Nota: V4 = versión 4, V5 = versión 5. El sombreado es para facilitar la lectura.

Para el análisis factorial confirmatorio se definió cada uno de los seis factores por sus cuatro ítems, y se utilizó el estimador robusto de mínimos cuadrados ponderados (WLSMV, por su denominación en inglés *Weighted Least Squares Mean and Variance Adjusted*) implementado en el programario Mplus 6.0 (Muhtén y Muthén, 1998-2010). Los índices de ajuste fueron muy similares en las dos versiones del cuestionario (χ^2 con 237 grados de libertad, entre 815,41 y 526,83; CFI de ,92; TLI entre ,90 y ,91; RMSEA entre ,07 y ,08).

El resto de análisis consistieron en la obtención de estadísticos descriptivos y de consistencia interna mediante el programario SPSS (v.15). Tal como puede verse en la Tabla 2, la consistencia interna de la mayoría de subescalas se mantuvo o mejoró en los datos del grupo que contestó la versión 5 del cuestionario, con la excepción de la subescala de regulación introyectada. Respecto al tipo de motivación, ambos grupos muestran resultados similares, con mayores promedios de motivación intrínseca y regulaciones de tipo autónomo y menores promedios de regulaciones de tipo controlado y amotivación.

Por lo que respecta a los estadísticos descriptivos de los ítems de las dos subescalas por las que nos interesamos especialmente en este trabajo, en la Tabla 3 se puede observar que los valores medios de los ítems en la versión 4 son relativamente bajos, indicando que los enunciados son falsos en promedio para este grupo. Lo mismo sucede en los promedios de la versión 5, con la notable excepción del ítem «porque me exijo continuar», cuyo promedio ha pasado al polo verdadero, con un valor más parecido al de las subescalas de regulación autónoma. La variabilidad de las respuestas es comparable entre ambas versiones, y respecto a la carga factorial, la excepción es el ítem «porque me exijo continuar» con una carga claramente inferior en la versión 5.

CONCLUSIONES

En este estudio se presenta de manera pormenorizada la evolución de dos ítems del cuestionario BRSQ durante el proceso de adaptación al idioma español. Dichos ítems son «I participate in my sport because people push me to play» que forma parte de la subescala de regulación externa; y «I participate in my sport because I feel obligated to continue» que forma parte de la subescala de regulación integrada.

La versión final propuesta para el cuestionario definitivo fue «practico este deporte porque los demás me exigen hacerlo» y «practico este deporte porque me siento obligada a continuar», por ser las adaptaciones que cumplen mejor con todas las condiciones, tanto las basadas en juicios de personas conocedoras, como las derivadas del análisis cuantitativo.

En particular, los datos cuantitativos apoyan la idea de que el concepto de exigencia por parte de los demás, está en línea con los de descontento, presión y satisfacción de los demás que se expresan en los distintos ítems de la subescala de regulación externa. En cambio, el concepto de autoexigencia no encaja con el resto de ítems de regulación introyectada que expresan ideas relacionadas con la vergüenza, el fracaso y la culpabilidad.

En el contexto de la práctica del deporte, las respuestas cuantitativas al ítem de autoexigencia están más cercanas a las dadas a los ítems que miden formas de regulación más autónoma. Sin embargo, conceptual y lingüísticamente hablando, un ítem con este contenido no tendría cabida en dichas subescalas, que están diseñadas para medir conceptos relacionados con la valoración de los beneficios que aporta el deporte y con la posibilidad que proporciona de expresar la manera de ser de uno mismo.

Creemos que este trabajo proporciona un buen ejemplo de los beneficios de triangular datos cualitativos y cuantitativos para proponer la versión más apropiada del cuestionario adaptado.

NOTA DE LOS AUTORES

Este trabajo se ha realizado, en parte, gracias a la subvención DEP 2010-15561 del Ministerio de Ciencia e Innovación. Los autores agradecen la colaboración de Fernando Azócar, Saül Alcaraz, Gustavo Korte, Alex Latinjak y Yago Ramis en las fases de adaptación lingüística y obtención de datos cuantitativos.

REFERENCIAS

Acquadro, C., Conway, K., Giroudet, Ch., y Mear, I. (2004). *Linguistic validation manual for patient-reported outcomes (PRO) instruments.* Lyon: Mapi Research Institute.

Deci, E. L., y Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. En R. K. Hambleton, P. F. Merenda and C. D. Spielberger (Eds.). *Adapting Psychological and Educational Tests for Cross-Cultural Assessment*. (pp.3-38). Mahwah, NJ: Lawrence Erlbaum.

Lonsdale, C., Hodge, K., y Rose, E.A. (2008). The behavioural regulation in sport questionnaire (BRSQ): Instrument development and initial validity evidence. *Journal of Sport and Exercise Psychology, 30*, 323-355.

Muthén, L. K., y Muthén, B. O. (1998-2010). *Mplus User's Guide*. (6th ed.). Los Angeles, CA: Muthén & Muthén.

Ramis, Y., Torregrosa, M., Viladrich, C., y Cruz, J. (2010). Adaptación y validación de la versión española de la Escala de Ansiedad Competitiva SAS-2 para deportistas de iniciación. *Psicothema, 22*, 1004-1009.

SPSS (Version 15.0) [Programario]. Sommers, NY: IBM.

Vallerrand, R. J. (2007). A hierarchical model of intrinsic and extrinsic motivation for sport and physical activity. En M.S. Hagger, y N.L.D. Chatzisarantis, (Eds.). *Intrinsic motivation and self-determination in exercise and sport*. (pp.255-279). Champaign, IL: Human Kinetics.

Viladrich, C., Torregrosa, M. y Cruz, J. (2011). Calidad psicométrica de la adaptación española del Cuestionario de Regulación Conductual en el Deporte. *Psicothema, 23,* 786-794.

UTILIZACIÓN DE LA ENTREVISTA COGNITIVA PARA OBTENER EVIDENCIAS SOBRE LOS PROCESOS DE RESPUESTA DE LOS INFORMANTES DIRECTOS-PROXY

Miguel Castillo y José Luís Padilla
Universidad de Granada
Correo electrónico: miguelcastillo@ugr.es

Resumen

La introducción de informantes proxy en los estudios por encuesta aporta ciertos beneficios como aumentar el tamaño de la muestra y la potencia del estudio (Campbell, Sloan & Kreiger, 2000). No obstante, también plantea algunas desventajas como el aumento del error de respuesta. Las respuestas distintas entre ambos informantes son el resultado de procesos de respuesta también distintos. El objetivo de este estudio fue explorar en qué fases del Modelo Pregunta-Respuesta (Tourangeau, 1984) aparecen las diferencias. Se realizaron 24 entrevistas cognitivas a 12 parejas de participantes usando preguntas provenientes del Cuestionario de Discapacidad de la Encuesta de Discapacidad, Autonomía Personal y Situaciones de Dependencia. Se utilizó el esquema de análisis de Willis, Schechter y Whitaker (1999) y el software AQUAD, versión 6.8.1.1. (Huber, 2008) para analizar los datos. Los resultados mostraron que ambos participantes ofrecían respuestas similares en un 67% de las ocasiones, con un índice de acuerdo K=.35. Las diferencias en los procesos de respuesta se encontraban principalmente en las fases de *Comprensión* y *Recuperación de la Información*. Los resultados indican que el primer paso hacia la confluencia de las respuestas parte de un mayor trabajo sobre la redacción y el contenido de las preguntas de los cuestionarios.

Actualmente, está aumentando el uso de informantes directos e informantes proxy en las encuestas nacionales. Un informante *proxy* es aquella persona que responde las preguntas de la encuesta poniéndose en el lugar de una tercera persona. (Magaziner, Bassett, Hebel & Gruber-Maldini, 1996). La introducción de estos informantes en los estudios por encuesta aporta beneficios, aumenta el tamaño de la muestra y la potencia del estudio (Campbell, Sloan & Kreiger, 2000), y reduce costes (White & Massey, 1981). No obstante, varios estudios han mostrado diferencias consistentes en los datos aportados por ambos (Bassett, Magaziner & Hebel, 1990; Epstein, Hall, Tognetti, Son & Conant, 1989; Mathiowetz & Groves, 1985; Rothman, Hedrick, Bulcroft, Hickarn & Rubenstein, 1991).

La *Eurostat Task Force* consensúa a qué casos restringir el uso de informants *proxy*: a) cuando se trata de reemplazar a aquellas personas que son incapaces de

responder a las preguntas debido a graves problemas de salud (p. ej. demencia, discapacidad física o mental severa, etc.), y b) cuando no es posible encuestar a ciertos grupos de personas por motivos legales, por ejemplo, menores de edad (Tafforeau, Lopez, Tolonen, Scheidt-Nave & Tinto, 2006).

Varios estudios han encontrado diferencias en la información ofrecida por un informante proxy de la que se obtendría a través de un informante directo, en relación a aspectos de salud observables y no observables (Bassett, Magaziner & Hebel, 1990; Epstein et al., 1989; Mathiowetz & Groves, 1985; Rothman et al., 1991). Grootendorst, Feeny y Furlong (1997) afirmaron que el estado emocional, el dolor y otras dimensiones de salud no directamente observables, son menos adecuadas para ser medidas utilizando informantes *proxy*.

Desde un punto de vista del Modelo del Proceso Pregunta-Respuesta de Tourangeau (1984), estas respuestas distintas serían el resultado de un proceso de respuesta también distinto. Basándonos en las fases del modelo (compresión, recuperación de la información, juicio y respuesta) encontraríamos diferencias en algunas de las fases entre ambos informantes. Dada esta situación es fundamental explorar en qué parte del proceso, el informante *proxy* introduce diferencias con respecto al informante directo. Para alcanzar este objetivo entrevistamos a parejas de informantes directos y *proxy* utilizando el método de la entrevista cognitiva con el fin de indagar en los procesos de respuesta. Este método de evaluación de preguntas de cuestionarios ha sido ampliamente aplicado como técnica de indagación en los procesos de respuesta (Castillo, Padilla, Gómez y Andrés, 2010; Collins, 2003; Willis, 2004).

MÉTODO

Participantes

El número de participantes en el estudio fue de 24 personas, 8 hombres y 16 mujeres, con edades comprendidas entre los 18 y los 65 años. Se balanceó el nivel educativo entre las categorías estudios primarios, estudios medios y estudios universitarios.

Se realizó la siguiente pregunta de reclutamiento: *«¿Alguna persona del hogar tiene alguna limitación para realizar las actividades que la gente habitualmente hace debido a un problema de salud?*». Se seleccionaron los participantes formando parejas en las que el miembro del hogar con una limitación asume el rol de informante directo mientras que otro miembro hace de informante proxy. La captación de los participantes fue realizada mediante el procedimiento de «bola de nieve».

Materiales

Se utilizó un protocolo de entrevista que incluía preguntas de un cuestionario de discapacidad y pruebas de indagación del proceso de respuesta. Las preguntas

del cuestionario habían sido desarrolladas a partir de la Clasificación Internacional del Funcionamiento, de la Discapacidad y de la Salud (CIF, 2001). A través de un comité de expertos se seleccionaron 11 preguntas del cuestionario de discapacidad.

Recogida de los datos

Todas las entrevistas se llevaron a cabo de forma individual por dos entrevistadores entrenados y experimentados en la realización de entrevistas cognitivas. Tras el consentimiento de los participantes las entrevistas se grabaron en audio-vídeo, y tuvieron lugar en un laboratorio cognitivo equipado de un sistema de grabación y digitalización en la Facultad de Psicología de la Universidad de Granada. Tras cada pregunta del cuestionario se aplicaron pruebas específicas para indagar en los procesos de respuesta.

Análisis de datos

Las 24 entrevistas cognitivas realizadas se analizaron basándonos en un esquema de codificación establecido por Willis, Schechter y Whitaker (1999). En el estudio se utilizó este esquema de codificación para identificar en qué fases del modelo del proceso pregunta-respuesta (Tourangeau, 1984) el informante proxy introducía diferencias.

El análisis de las entrevistas cognitivas se estructuró por parejas, comparando los informes verbales de cada miembro de la pareja ante cada pregunta del cuestionario. Para el análisis de los datos de las entrevistas se utilizó el software AQUAD, versión 6.8.1.1. (Huber, 2008).

RESULTADOS

Convergencia de las respuestas a las preguntas del cuestionario

La distribución de las 132 respuestas obtenidas aparece representada en la Tabla 1. Dicha tabla muestra la frecuencia de las respuestas coincidentes (Sí/Sí; No/No), las no coincidentes (Sí/No; No/Sí), los datos perdidos y las respuestas con distinta graduación. Esta última categoría hace referencia a aquellas respuestas en las que ambos participantes responden «Sí» pero con distinta graduación. Por ejemplo, «Sí, estoy gravemente limitado» / «Sí, mi marido esta limitado, pero no gravemente».

Las respuestas de los informantes directos e informantes *proxy* coincidieron solamente en un 67.4%, mientras que ofrecieron respuestas distintas en un 22.7%. El índice de acuerdo *Kappa* de Cohen (1960) calculado sobre las respuestas de los participantes fue $K = .35$. Este valor muestra un nivel de acuerdo bajo entre ambos informantes.

Tabla 1. Tipos y Frecuencias de Respuestas

Respuestas Informante Proxy	Respuestas Informante Directo	
	Sí	No
Sí	13 (9.8%)	9 (6.8%)
No	18 (13.6%)	76 (57.6%)
Distinta Graduación	3 (2.3%)	
Perdidos	13 (9.9%)	
Total	132 (100%)	

Convergencia de los procesos de respuesta

En la Tabla 2, aparecen las respuestas de cada par de informantes a las preguntas del cuestionario y las veces que los analistas identificaron una diferencia en el proceso.

Tabla 2. Diferencias en el Proceso Pregunta-Respuesta

Respuesta: Proxy / Directo	Frecuencia	Diferencias en Proceso Respuesta
Sí / Sí	13	4
No / No	76	5
Sí / No	9	9
No / Sí	18	18
Sí / Sí (diferente graduación)	3	3
Perdidos	13	0
Total	132	39

Con respecto a las respuestas coincidentes, pese a que el informante proxy está ofreciendo la misma respuesta que el directo (Sí/Sí, No/No) no siempre coinciden sus procesos de respuesta.

Una vez que la respuesta del informante proxy es distinta (Sí/No, No/Sí, Sí/Sí distinta graduación) siempre existe una diferencia subyacente en los respectivos procesos de respuesta.

Identificación de las diferencias en los procesos de respuesta

Los analistas identificaron la mayoría de las diferencias en la fase de *Comprensión* y *Recuperación de Información* y, seguidamente, en la fase de *Juicio*. Los analistas no identificaron ninguna diferencia en la fase de *Respuesta*. El Informe Verbal 1 proviene de una pareja de participantes en la que el informante directo tiene una limitación visual.

Informe Verbal 1

P. ¿Debido a problemas de naturaleza cognitiva o intelectual, *usted / alguna persona del hogar* tiene dificultad importante para utilizar intencionadamente los sentidos? Por ejemplo, prestar atención con la mirada, mantener la atención con el oído, etc.

- **B1_05_F_50 (Directo)**: «*Sí, sí. Pongo más atención»*
- Entrevistador: «*Explícame lo que quieres decir»*
- **B1_05_F_50 (Directo)**: «*Una persona que se enfrenta a esta situación, como por ejemplo en mi caso, sí pones muchísimo el sentido, sobre todo el oído. Muchísimo. Yo creo que muchísimo más que antes, sobre todo memorizar. Yo lo que hago más ahora es memorizar; entonces oigo más y memorizo más. Y quizás a la hora de ver, memorizo más o me sitúo más en el sitio que estoy; y si veo una cosa que es verde me sitúo más por el sentido o por lo que he podido escuchar, pero sobre todo por el sentido, por la orientación del sentido. Y manejas mucho la memoria y el oído, escuchas mucho más que antes, mucho más»*
- **A1_05_F_30 (Proxy)**: «*Sí, mi madre»*
- Entrevistador: «*¿Qué significa esta pregunta para ti, por qué ha contestado así?»*
- **A1_05_F_30 (Proxy)**: «*Me he fijado sobre todo en la parte que me has aclarado cuando has dicho por ejemplo fijar la mirada atentamente y demás, porque no tengo muy claro que la alteración que tiene mi madre sea de naturaleza cognitiva o intelectual y que sea solamente física. Entonces por eso me he centrado en el ejemplo»*
- Entrevistador: «*Es decir, te has centrado en el ejemplo de prestar atención con la mirada*»
- **A1_05_F_30 (Proxy)**: «*Sí. A lo mejor si no hubieras dicho lo de prestar atención con la mirada, te hubiera contestado nada porque no lo hubiera sabido responder*»

Las diferencias se encuentran en las fases de *Compresión* y *Recuperación de la información*. *Comprensión*: El informante directo no tiene en cuenta la especificación «una dificultad importante por problemas de naturaleza cognitiva e intelectual», mientras que el informante proxy sí. Además, el informante directo fija su interpretación en los conceptos «atención y, mantener la atención con el oído»; en cambio el informante *proxy* se centra en el concepto «prestar atención con la mirada». *Recuperación de la información*: el informante directo tiene acceso y recupera información interna en la que procura memorizar más, orientarse, prestar más atención con el oído, etc. Mientras que el informante *proxy* recupera información externa en la que su madre tiene dificultades para prestar atención con la mirada.

El Informe Verbal 2 proviene de un participante directo con una grave limitación visual y su mujer como participante *proxy*.

Informe Verbal 2

P. *Usted / Alguna persona de su hogar*, ¿se ha visto limitado/a para realizar las actividades que la gente habitualmente hace, debido a un problema de salud? La limitación debe durar más de un año.

- **F1_03_M_62 (Directo)**: «*Sí, pero no gravemente limitado»*
- Entrevistador: «*¿Qué entiende por actividades de la vida cotidiana?»*
- **F1_03_M_62 (Directo)**: *«…actividades de la vida cotidiana son, levantarse, desayunar, ir al servicio, vestirse...esas puedo hacerlas, pero hay otras actividades de la vida cotidiana como es salir a la calle, cruzar la calle, ir a la tienda en estas sí que necesito ayuda*»

- **E1_03_F_51 (Proxy)**: «*Sí, gravemente limitado»*
- Entrevistador: «*¿Qué entiende por actividades de la vida cotidiana?»*
- **E1_03_F_51 (Proxy)**: «*…por ejemplo, levantarse por la mañana, ir al servicio, asearse sin que nadie te diga dónde está el cepillo o la pasta de dientes, si no encuentra el jabón que nadie te tenga que decir dónde está...no puede salir solo a la calle, tiene que ir con alguien...»*

La diferencia radica en la fase de *Recuperación de información*. El informante directo recupera dos tipos de actividades: una que podríamos denominar de «puertas hacia adentro» como levantarse, desayunar, ir al servicio, vestirse, etc., y otras actividades que podríamos denominar de «puertas hacia fuera» como salir a la calle, cruzar la calle, ir a la tienda, etc. Por otro lado, el informante *proxy* recupera información de todas las actividades en las que su marido tiene problemas, sin diferenciar entre ellas.

DISCUSIÓN

El objetivo de este estudio fue explorar las diferencias en los procesos de respuesta que lleva a cabo el informantes proxy con respecto al informante directo. Los resultados del estudio han mostrado, en primer lugar, el bajo nivel de acuerdo entre las respuestas de ambos informantes, con un índice *Kappa* de .35. En segundo lugar, a pesar de que la respuesta del informante proxy coincida con el directo, los procesos de respuesta subyacentes en ocasiones son distintos. En tercer lugar, una vez que ambos informantes han ofrecido una respuesta distinta existe una diferencia subyacente en el proceso de pregunta-respuesta. La mayoría de estas diferencias subyacentes se encuentran en la fase de *compresión* y *recuperación de información* del modelo del proceso de pregunta-respuesta (Tourangeau, 1984). En la bibliografía se establecía que, las diferencias en las respuestas se deben a la distinta información disponible para realizar juicios y a las distintas estrategias de formación de dichos juicios (Bickart et al., 1994; Kuiper & Rogers, 1979; Schwarz & Wellens, 1997; Sudman, et al., 1996). El presente estudio ha mostrado que además de las diferencias en la disponibilidad de la información hay que tener en

cuenta las diferentes interpretaciones que realizan los participantes de las preguntas del cuestionario.

En este estudio se ha demostrado, una vez más, cómo el uso de la entrevista cognitiva nos permite saber cómo están interpretando y respondiendo a las preguntas de las encuestas los participantes e identificar los tipos de error que se están cometiendo (DeMaio & Rothgeb, 1996; Harris-Kojetin, Fowler, & Brown, 1999; Nápoles-Springer, Santoyo-Olsson, O'Brien, & Stewart, 2006).

REFERENCIAS

Bassett, S.S., Magaziner, J. & Hebel, J.R. (1990) Reliability of proxy response on mental health indices for aged, community-dwelling women. *Psychology and Aging, 5*, 127-132.

Bickart, B., Menon, G., Schwarz, N. & Blair, J. (1994). The use of anchoring strategies in constructing proxy reports of attitudes. *International Journal of Public Opinion Research, 6*, 375-379.

Campbell, P.T., Sloan, M. & Kreiger, N. (2000). Utility of Proxy versus Index Respondent Information in a Population-Based Case-Control Study of Rapidly Fatal Cancers. *Annals of Epidemiology, 17*, 253-257.

Castillo, M., Padilla, J.L., Gómez, J. & Andrés, A. (2010). A productivity map of cognitive pretest methods for improving survey question. *Psicothema, 22* (3), 475-481.

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. Educational and Pychological Measurement, 20, 37-46.

Collins, D. (2003). Prestesting survey instruments: An overview of cognitive methods. *Quality of Life Research, 12*, 229-238.

DeMaio, T.J. & Rothgeb J.M. (1996). Cognitive interviewing techniques: in the lab and in the field. En N. Schwarz & S. Sudman (Eds.), *Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francsico: Jossey-Bass, 177–196.

Epstein, A.M., Hall, J.A., Tognetti, J., Son, L.H. & Conant, L. (1989). Using proxies to evaluate quality of life. *Medical Care, 27,* 91-98.

Grootendorst, P.V., Feeny, D.H. & Furlong, W. (1997). Does it matter whom and how you ask? Inter- and intra-rater agreement in the Ontario Health Survey. *Journal of Clinical Epidemiology, 50*, 127-135.

Harris-Kojetin, L.D., Fowler, F.J. & Brown, J.A. (1999). The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. Consumer Assessment of Health Plans Study. *Medical Care, 37*, 10–21.

Huber, G.L. (2008). AQUAD El Programa para Analizar Datos Cualitativos (versión 6.8.1.1). Universidad de Tübingen, Alemania: Ingeborg Huber Verlag.

Kuiper, N.A. & Rogers, T.B. (1979). Encoding of personal information: self-other differences. *Journal of Personality and Social Psychology, 37*, 499-514.

Magaziner, J., Bassett, S.S., Hebel, J.R. & Gruber-Maldini, A. (1996). Use of Proxies to Measure Health and Functional Status in Epidemiologic Studies of Community-dwelling Women Aged 65 Years and Older. *American Journal of Epidemiology, 143*, 283-292.

Mathiowetz, N.A. & Groves, R. M. (1985). The effects of respondent rules on health survey reports. *American Journal of Public Health, 75,* 311-316.

Nápoles-Springer, A.M., Santoyo-Olsson, J., O'Brien, H. & Stewart, A.L. (2006). Using cognitive interviews to develop surveys in diverse populations. *Medical Care, 44*, 21-30.

Rothman, M.L., Hedrick, S.C., Bulcroft, K.A., Hickarn, D.H. & Rubenstein, L.Z. (1991). The validity of proxy-generated scores as measures of patient health status. *Medical Care, 29,*115-124.

Schwarz, N. & Wellens, T. (1997). Cognitive dynamics of proxy responding: the diverging perspective of actors and observers. *Journal of Official Statistics, 13*, 159-179.

Sudman, S., Bradburn, N.M. & Schwarz, N. (1996). *Thinking about answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass.

Tafforeau, J., Lopez, M., Tolonen, A., Scheidt-Nave, C. & Tinto, A. (2006). *Guidelines for the development and criteria for the adoption of Health Survey instruments*. Eurostat Working Papers and Studies.

Tourangeau, R. (1984). Cognitive science and survey methods: a cognitive perspective. En: T. Jabine, M. Straf, J. Tanur & R. Tourangeau (Eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between the Disciplines* (pp. 73-100). Washington, DC: National Academy Press.

White, A.A. & Massey, J.T. (1981). Selective reduction of proxy response bias in a household interview survey. Proceedings of the Social Statistics Section, 211-216. American Statistical Association: Washington, DC.

Willis, G.B. (2004). Cognitive Interviewing Revisited: A Useful Technique, in Theory? En S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questions* (pp. 23-43). Nueva Jersey: John Wiley & Sons.

Willis, G.B., Schechter, S. & Whitaker, K. (1999). A comparison of cognitive interviewing, expert review, and behavior coding: What do the tell us? *Proceedings de la Section on Survey Research Methods, American Statistical Association,* 28-37.

FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

Coordinadoras: Juana Gómez-Benito[1] y M. Dolores Hidalgo[2]
[1] *Universidad de Barcelona*
[2] *Universidad de Murcia*

Los estudios de Funcionamiento Diferencial de los ítems (Diferential Item Functioning, DIF) constituyen una de las líneas de investigación psicométrica que más interés y aplicabilidad ha suscitado en las últimas décadas. Este simposio pretende hacerse eco de esta preocupación por garantizar la equivalencia de los instrumentos de medida en los diversos grupos en los que pueden aplicarse. Las aportaciones se centran en el estudio de la eficacia de distintos procedimientos de detección de DIF que se insertan en los avances metodológicos debatidos en los últimos años, como la prueba de Breslow-Day, la regresión logística ordinal, o el análisis factorial confirmatorio con estructuras de medias y covarianzas, entre otros. Al mismo tiempo abordan aspectos novedosos de evidente utilidad práctica, como la disociación entre DIF paralelo y DIF uniforme en función del modelo utilizado, el efecto que producen la cantidad de DIF en el test al realizar comparaciones entre grupos, o la posibilidad de interacciones cuando los grupos de comparación difieren en varias características relevantes. El simposio aborda en conjunto cuestiones metodológicas y sustantivas, implicando técnicas diversas y contextos de aplicación diferentes, en aras del constante incremento de calidad de los instrumentos de medida psicológicos.

PALABRAS CLAVE: DIF, Invarianza, Sesgo.

DIF PARALELO VS. DIF UNIFORME

María Ester Aguerri[1], Pedro Prieto[2], María Silvia Galibert[1] y Horacio F. Attorresi[1]

[1] Universidad de Buenos Aires

[2] Universidad de La Laguna

Correo electrónico: *maguerri@psi.uba.ar*

Resumen

La evaluación de métodos de detección del funcionamiento diferencial del ítem (DIF) se realiza sobre la base de estudios de simulación. Entre los modelos más utilizados para la generación de las respuestas a ítems dicotómicos está el modelo logístico de tres parámetros (ML3P). Los distintos tipos de DIF se generan frecuentemente sobre la base de diferentes situaciones para las curvas características del ítem (CCIs). Hanson (1998) sostuvo que habría que distinguir entre DIF Paralelo (las CCIs son paralelas), DIF Uniforme (la función de los *Odds Ratio* es una constante diferente de la unidad) y DIF Unidireccional (las CCIs no se cortan). Mostró que si bien todo DIF Paralelo es Unidireccional y todo DIF Uniforme es Unidireccional, no todo DIF Paralelo es Uniforme ni todo DIF Uniforme es Paralelo. En este trabajo se precisan las relaciones que verifican los parámetros del ítem cuando existe DIF Uniforme, siguiendo la línea planteada por Hanson. Los resultados obtenidos consolidan la disociación entre DIF Paralelo y DIF Uniforme para el modelo ML3P. Esto es, si el modelo es ML3P: cuando las CCIs son paralelas el DIF es no Uniforme y cuando el DIF es Uniforme las CCIs no son paralelas.

En el marco de la Teoría de Respuesta al Ítem (TRI) la ausencia o presencia de DIF está dada por la coincidencia o no de las curvas características del ítem (CCIs), esta concepción es aceptada unánimemente. No ocurre lo mismo con la clasificación del DIF. La mayor parte de los estudios reconoce la diferencia entre DIF Uniforme y DIF no Uniforme siguiendo la clasificación de Mellenbergh (1982). Así, que un ítem presente DIF Uniforme significa que hay una diferencia constante en el rendimiento en el ítem que favorece a uno de los grupos a lo largo de los niveles de habilidad. Cuando tal diferencia no es constante, el DIF es no Uniforme. Swaminathan y Rogers (1990), en términos de la TRI, afirmaron que el DIF no Uniforme está indicado por CCIs no paralelas. Ackerman (1992) sostuvo que un ítem presenta DIF Uniforme si sus CCIs difieren sólo por una traslación horizontal, o sea son paralelas pero no coincidentes. Ackerman hizo notar que para los modelos logísticos de dos y tres parámetros (ML2P y ML3P) las CCIs pueden ser no paralelas dando lugar a DIF no Uniforme. Mazor, Clauser y Hambleton (1994)

afirmaron que el DIF no Uniforme puede pensarse en términos de la TRI como una diferencia en los parámetros de discriminación y reservaron, de hecho, la denominación de DIF Uniforme para el caso en que las CCIs sólo difieren en el parámetro de dificultad. Insistieron en esta línea Clauser y Mazor (1998), al decir que un ítem presenta DIF Uniforme si las CCIs difieren sólo en el parámetro de dificultad, y no Uniforme en el resto de los casos. Por otra parte Zumbo (1999) afirmó que cuando hay DIF Uniforme las CCIs no se cortan y cuando hay DIF no Uniforme las CCIs sí se cortan. Esta postura también fue considerada en trabajos donde se estudió la potencia de métodos de detección del DIF no Uniforme, pues consideraron para ello exclusivamente ítems cuyas CCIs son no paralelas. Así Mazor, Clauser y Hambleton (1994) evaluaron al procedimiento de Mantel-Haenszel modificado, Swaminathan y Rogers (1990) presentaron el estudio del DIF mediante la regresión logística y Penfield (2003) aplicó la prueba de la tendencia en la heterogeneidad de los *Odds Ratio* de Breslow-Day. Penfield y Camilli (2007) y Osterlind y Everson (2009) mencionaron que la distinción entre DIF Uniforme y no Uniforme tiene importancia crítica en los estudios del DIF, pues los métodos estadísticos para su detección son diferentes. No obstante en ninguno de los dos trabajos se explicitaron las relaciones verificadas en cada situación de DIF por los parámetros del ítem. Osterlind y Everson (2009) ilustraron los distintos tipos de DIF con CCIs correspondientes al ML3P: al DIF no Uniforme con dos CCIs que se cortan y al DIF Uniforme con dos CCIs que no se cortan, sin mencionar los valores de los parámetros de tales ítems. Es decir, distintos estudios han caracterizado al DIF Uniforme por la posición relativa de las dos CCIs en el plano, bien asociándolo con CCIs paralelas o bien sin especificar las relaciones que verifican los parámetros del ítem en los grupos intervinientes.

Por su parte, Hanson (1998) sostuvo que habría que distinguir entre DIF Paralelo (las CCIs son paralelas), DIF Uniforme y DIF Unidireccional (las CCIs no se cortan). Mostró que: (1) todo DIF Paralelo es Unidireccional, (2) todo DIF Uniforme es Unidireccional, (3) no todo DIF Paralelo es Uniforme, y (4) no todo DIF Uniforme es Paralelo. Estas dos últimas afirmaciones sobre la base de contraejemplos y sin caracterizar de manera completa al DIF Uniforme.

El objetivo del presente trabajo es diferenciar al DIF Paralelo del DIF Uniforme, y precisar las relaciones que verifican los parámetros del ítem en los grupos involucrados cuando existe DIF Uniforme, siguiendo la línea planteada por Hanson (1998).

DESARROLLO

Si bien Hanson (1998) trabajó con el ML3P con la constante de escalamiento $D = 1$, en el presente trabajo lo haremos con $D = 1.7$ que es el usual en la mayor parte de los estudios del DIF.

Consideremos la función de respuesta a un ítem dicotómico según el ML3P en los Grupos 1 y 2:

$$P_1(w) = c_1 + (1 - c_1)\frac{1}{1 + e^{-1.7.a_1.(w - b_1)}},\ \forall w \in \Re \text{ y}$$

$$P_2(w) = c_2 + (1 - c_2)\frac{1}{1 + e^{-1.7.a_2.(w - b_2)}},\ \forall w \in \Re.$$

Las constantes a_i, b_i y c_i con i = 1, 2 son, respectivamente, los parámetros de discriminación, dificultad y aciertos por azar del ítem en el Grupo 1 y en el Grupo 2, y son tales que a_1, a_2 ε $\Re^+$, b_1 y b_2 ε $\Re$ y c_1 y c_2 ε $\Re$ con $0 \leq c_1, c_2 < 1$.

Hanson (1998) denominó DIF Paralelo al que se corresponde con CCIs paralelas.

Para definir y caracterizar al DIF Uniforme, Hanson (1998) recurrió a la función de los *Odds Ratio,* θ(*w*), medida de asociación originalmente formulada por Agresti (1990). La expresión de θ(*w*) para el caso de ítems dicotómicos respondidos por dos grupos es:

$$\theta(w) = \frac{P_1(w).[1 - P_2(w)]}{P_2(w).[1 - P_1(w)]},\ \forall w \in \Re.$$

Considerando el ML3P resulta:

$$\theta(w) = \frac{(1 - c_2).\left[c_1 + e^{1.7a_1(w - b_1)}\right]}{(1 - c_1).\left[c_2 + e^{1.7a_2(w - b_2)}\right]},\quad \forall w \in \Re. \qquad [1]$$

Hanson (1998) afirmó que un ítem presenta DIF Uniforme si existe $\alpha \in \Re^+$, $\alpha \neq 1$, tal que:

$$\theta(w) = \alpha,\ \forall w \in \Re. \qquad [2]$$

Es decir, el DIF es Uniforme si la función de los *Odds Ratio*, θ(*w*), es una función constante, distinta de la unidad, y no Uniforme si dicha función no es constante. Hanson (1998, p. 246) sostuvo que la definición [2] de DIF Uniforme es consistente con la definición de Mellenbergh (1982).

Hanson (1998) mostró que si existe DIF Uniforme los parámetros de aciertos por azar de las CCIs satisfacen la siguiente relación:

$$c_2 = \frac{c_1}{\alpha + (1 - \alpha)c_1},\ \alpha \in \Re^+,\ \alpha \neq 1 \text{ y } 0 \leq c_1, c_2 < 1 \qquad [3]$$

Nos preguntamos si el parámetro de aciertos por azar puede ser el mismo en los dos grupos ($c_1 = c_2 = c$) cuando hay DIF Uniforme. Si así fuera, la expresión [3] resulta igual a:

$$c = \frac{c}{\alpha + (1 - \alpha)c},\quad \alpha \in \Re^+,\ \alpha \neq 1 \text{ y } 0 \leq c < 1$$

Operando se obtiene que $c = 0$, o sea, existe DIF Uniforme con el mismo parámetro de aciertos por azar en los grupos si el modelo logístico ajustado es el de uno o dos parámetros (ML1P y ML2P).

Ahora bien, la función de los *Odds Ratio* para el ML2P tiene la siguiente expresión:

$$\theta(w) = \frac{e^{1.7a_1(w-b_1)}}{e^{1.7a_2(w-b_2)}}, \forall w \in \Re$$

Para que exista DIF Uniforme debe existir una constante α, real positiva y distinta de la unidad tal que $\theta(w) = \alpha$, $\forall w \in \Re$. Operando para que esto ocurra resulta que debe ser $a_1 = a_2 = a$, y $b_1 \neq b_2$, es decir que las CCIs son paralelas.

De manera análoga se obtiene en el ML1P que si el DIF es Uniforme las CCIs son paralelas.

Elegido el ML3P, ¿qué relación verifican los parámetros de las CCIs para que el DIF sea Uniforme? Sabemos que c_1 y c_2 están vinculados por [3] pero, ¿qué relación deben verificar, respectivamente, a_1 y a_2, y b_1 y b_2?

El DIF es Uniforme en tanto exista $\alpha \in \Re^+$, $\alpha \neq 1$, tal que en [1] resulte:

$$\theta(w) = \alpha = \frac{(1-c_2)\left[c_1 + e^{1.7a_1(w-b_1)}\right]}{(1-c_1)\left[c_2 + e^{1.7a_2(w-b_2)}\right]}, \forall w \in \Re.$$

De donde debe ser:

$$\alpha(1-c_1)\left[c_2 + e^{1.7a_2(w-b_2)}\right] = (1-c_2)\left[c_1 + e^{1.7a_1(w-b_1)}\right], \forall w \in \Re.$$

Operando:

$$\alpha.(1-c_1).c_2 - (1-c_2)c_1 = (1-c_2)e^{1.7a_1(w-b_1)} - \alpha(1-c_1)e^{1.7a_2(w-b_2)}, \forall w \in \Re.$$

Por [3] el primer miembro es nulo, luego:

$$(1-c_2)e^{1.7a_1(w-b_1)} = \alpha(1-c_1)e^{1.7a_2(w-b_2)}, \forall w \in \Re.$$

Operando y despejando, resulta que:

$$\frac{(1-c_2)}{\alpha(1-c_1)} e^{1.7(a_2b_2 - a_1b_1)} = e^{1.7(a_2-a_2)w}, \forall w \in \Re.$$

Por tanto debe ser $a_1 = a_2 = a$, luego:

$$e^{1.7a(b_2-b_1)} = \frac{\alpha.(1-c_1)}{(1-c_2)}$$

Por [3] resulta que:

$$e^{1.7a(b_2-b_1)} = \frac{c_1}{c_2}$$

Tomando logaritmo y despejando resulta que:

$$b_2 = b_1 + \frac{1}{1.7a} log \frac{c_1}{c_2}$$

Luego en el ML3P el DIF es Uniforme si:

$$a_1 = a_2 = a,\ b_2 = b_1 + \frac{1}{1.7a} log \frac{c_1}{c_2} \text{ y } c_2 = \frac{c_1}{\alpha + (1-\alpha)c_1} \qquad [4]$$

y estas CCIs no son paralelas porque $c_1 \neq c_2$ ($\alpha \neq 1$).

Veamos la expresión de la función de los *Odds Ratio* para un ítem con DIF Uniforme modelizado con el ML3P conocido el parámetro de aciertos por azar del ítem en los grupos intervinientes. Reemplazando en [1] por las relaciones verificadas por los parámetros de dificultad y discriminación señaladas en [4], resulta $\forall w \in \Re$.

$$\theta(w) = \frac{(1-c_2)\left[c_1 + e^{1.7a(w-b_1)}\right]}{(1-c_1).\left[c_2 + e^{1.7a\left(w-b_1-\frac{1}{1.7a}log\frac{c_1}{c_2}\right)}\right]}$$

Luego,

$$\theta(w) = \frac{(1-c_2)\left[c_1 + e^{1.7a(w-b_1)}\right]}{(1-c_1).\left[c_2 + \frac{c_2}{c_1}e^{1.7a(w-b_1)}\right]}$$

Y operando resulta:

$$\theta(w) = \frac{(1-c_2).c_1}{(1-c_1).c_2}, \text{ que es constante.}$$

Por ser $0 \leq c_1, c_2 < 1$ y $c_1 \neq c_2$ resulta $\theta(w) > 0$ y $\theta(w) \neq 1$.

El valor constante que corresponde a la función de los *Odds Ratio* en el caso del DIF Uniforme da cuenta de la magnitud del DIF.

A continuación se presentan ejemplos de ítems con DIF Uniforme en el ML3P en los que el Grupo 1 es el Grupo de Referencia (GR) y el Grupo 2 es el Grupo Focal (GF).

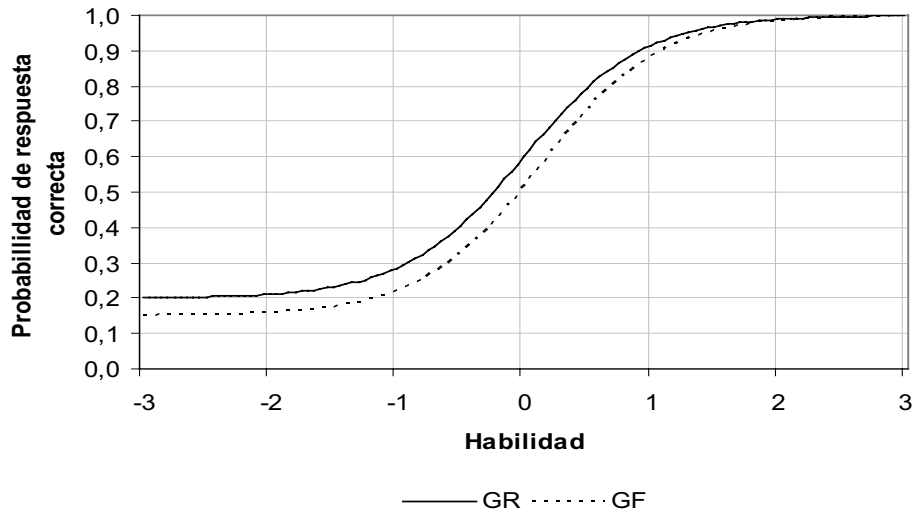


Gráfico 1. Curvas características de un ítem con DIF Uniforme, $\theta(w) = 1.42$ para todo $w \in \Re$, con parámetros a = 1.25, b = 0 y c = 0.2 en GR y parámetros a = 1.25, b = 0.1354 y c = 0.15 en GF

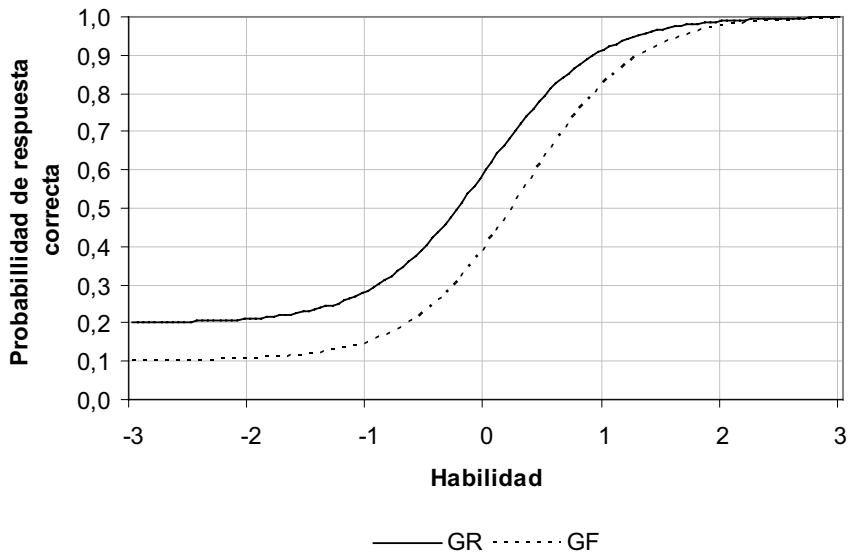


Gráfico 2. Curvas características de un ítem con DIF Uniforme, $\theta(w) = 2.25$ para todo $w \in \Re$, con parámetros a = 1.25, b = 0 y c = 0.2 en GR y parámetros a = 1.25, b = 0.3262 y c = 0.1 en GF

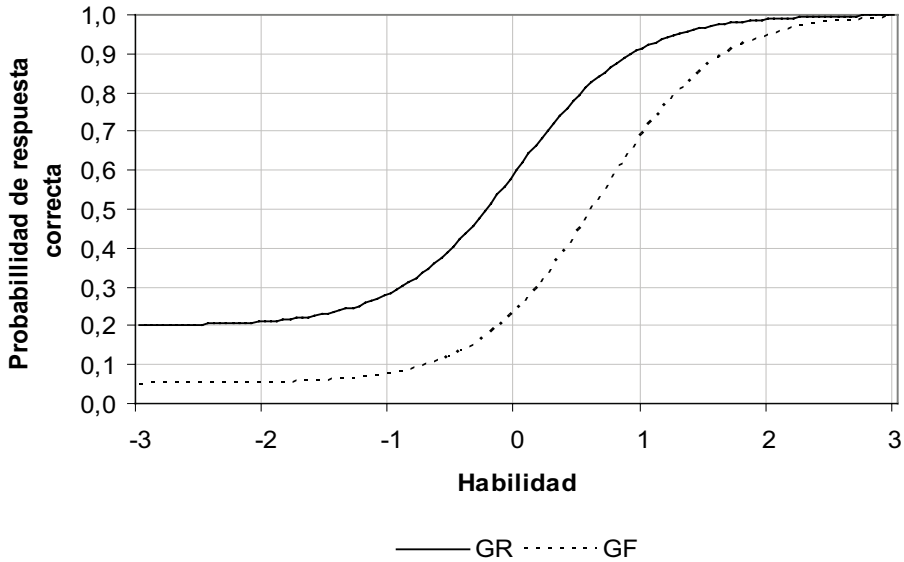


Gráfico 3. Curvas características de un ítem con DIF Uniforme, $\theta(w) = 4.75$ para todo $w \in \Re$, con parámetros a = 1.25, b = 0 y c = 0.2 en GR y parámetros a = 1.25, b = 0.6524 y c = 0.05 en GF

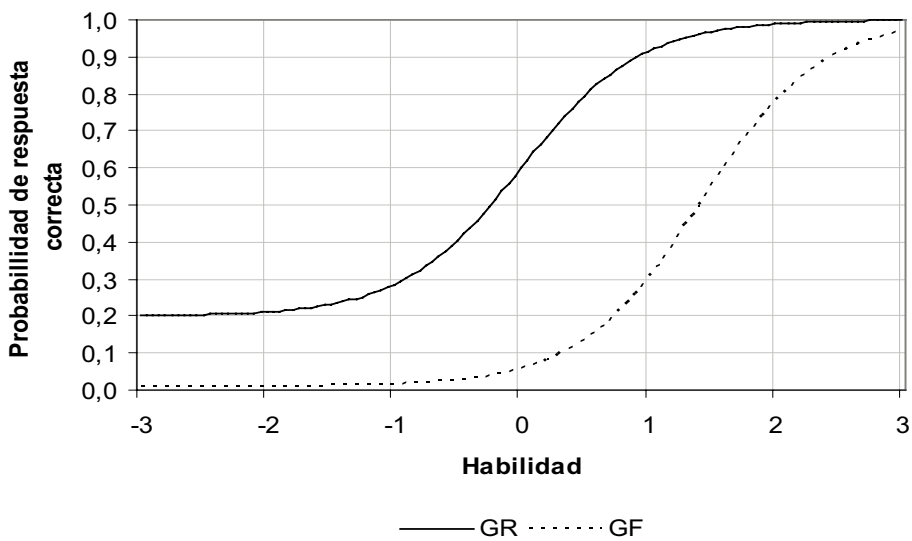


Gráfico 4. Curvas características de un ítem con DIF Uniforme, $\theta(w) = 24.75$ para todo $w \in \Re$, con parámetros a = 1.25, b = 0 y c = 0.2 en GR y parámetros a = 1.25, b = 1.4078 y c = 0.01 en GF

Veamos en qué situaciones el DIF Paralelo es Uniforme. El DIF es Paralelo cuando las CCIs son paralelas, esto es cuando sólo difieren en el parámetro de dificultad.

En ML3P las CCIs son paralelas si se verifica que $c_1 = c_2 = c \neq 0$, $a_1 = a_2 = a$, y $b_1 \neq b_2$. Por tanto la función de los *Odds Ratio* resulta:

$$\theta(w) = \frac{c + e^{1.7a(w-b_1)}}{c + e^{1.7a(w-b_2)}}, \ \forall w \in \Re.$$

La expresión anterior depende de w, y por tanto no es constante. Por lo tanto en el modelo logístico de tres parámetros con igual parámetro de acierto por azar para todos los ítems (ML3P-*c*), el DIF Paralelo no es Uniforme.

Para el ML2P, si hay DIF Paralelo, debe ser $a_1 = a_2 = a$ y $b_1 \neq b_2$, luego en [1] resulta $\theta(w) = e^{1.7a(b_2-b_1)}$ que es constante, y distinta de 1, $\forall w \in \Re$. Luego, hay DIF Uniforme. Lo mismo ocurre para ML1P.

CONCLUSIÓN

Lo expuesto consolida la disociación entre DIF Paralelo y DIF Uniforme cuando el modelo es el ML3P. Como se ha visto, para el ML3P si las CCIs son paralelas el DIF es no Uniforme y si existe DIF Uniforme las CCIs son no paralelas. Sí se mantiene la equivalencia entre DIF Paralelo y DIF Uniforme cuando el modelo es ML2P o ML1P.

Las investigaciones que evaluaron métodos de detección del DIF no Uniforme, y consideraron el ML3P-*c*, han dejado un vacío si no consideraron la presencia de DIF no Uniforme Paralelo. Es así que en futuros estudios de simulación que usen el modelo ML3P-*c*, se ha de tener presente toda situación de DIF es de DIF no Uniforme.

En el caso de estudiar la potencia de métodos aptos para identificar el DIF Uniforme, se los debería evaluar sobre ítems con tal tipo de DIF y por tanto modelizados con ML1P, ML2P, o bien ML3P, con parámetros del ítem que verifiquen las relaciones [4].

Cuando métodos diseñados para detectar DIF Uniforme señalan con DIF a ítems con DIF Paralelo modelizados con el ML3P-*c*, más que rechazar la hipótesis nula de ausencia de DIF podrían estar indicando el incumplimiento de los supuestos necesarios para su aplicación.

La caracterización de los ítems con DIF Uniforme en el ML3P, al menos respecto de la variable ideal, permite incorporar una nueva familia de ítems sobre los cuales los métodos de detección del DIF tendrán que ser evaluados.

Por otra parte, la función de los *Odds Ratio* como medida del DIF para el caso de DIF Uniforme en el ML3P completa, en parte, el vacío observado para las medidas del área de Raju (1988).

REFERENCIAS

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley.

Clauser, B.E., & Mazor, K. M. (1998). Using statistical procedures to identify differential functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23,* 244-253.

Mazor, K., Clauser, B., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using variation of the Mantel-Haenszel procedure. *Journal of Educational Measurement, 54*, 284-291.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-108.

Osterlind, S. J. & Everson, H.T. (2009). *Differential Item Functioning* (2nd edition). Thousand Oaks, California: Sage Publications, Inc.

Penfield, R. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *The Alberta Journal of Educational Research, 49*, 231-243.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. En C. R. Rao & S. Sinharay (Eds.) *Handbook of Statistics* (Vol. 26, pp. 125-167). New York: Elsevier.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 284-291.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007).Three generations of DIF analysis: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

EVALUACIÓN DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ITEMS EN ESCALAS ACTITUDINALES DEL PISA: UNA APLICACIÓN PARA ÍTEMS POLITÓMICOS DEL ESTADÍSTICO MANTEL-HAENZSEL Y LA REGRESIÓN LOGÍSTICA ORDINAL

Isabel Benítez Baena[1], José Luis Padilla[1], Mª Dolores Hidalgo Montesinos[2] y Stephen G. Sireci[3]

[1] Universidad de Granada

[2] Universidad de Murcia

[3] Universidad de Massachusetts Amherst

Correo electrónico: ibenitez@ugr.es

Resumen

En las últimas décadas, el Funcionamiento Diferencial de los Items (DIF por sus siglas en inglés) ha centrado la atención de profesionales implicados en evaluaciones internacionales y transculturales. La utilidad de evaluar el DIF en este tipo de estudios radica en la posibilidad de obtener evidencias sobre el nivel de equivalencia entre distintas versiones lingüísticas o culturales. El objetivo de este estudio es ilustrar la utilización de dos procedimientos, el estadístico Mantel-Haenzsel y la Regresión Logística Ordinal (RLO), para detectar DIF en ítems politómicos actitudinales. Para ello, se utilizaron las respuestas de 17405 estudiantes, procedentes de España y de Estados Unidos, a siete de las escalas actitudinales incluidas en el Programa de evaluación internacional de estudiantes (PISA, OECD, 2006). En estas muestras extraídas de la base de datos de la OECD se evaluó el DIF aplicando ambos procedimientos y posteriormente se compararon los resultados utilizando la log-odd ratio como medida del tamaño del efecto y el índice DELTA de la Educational Testing Service como criterio para determinar la cantidad de DIF. Los resultados mostraron ocho ítems identificados con DIF «alto» por ambos procedimientos. También aparecieron algunas diferencias que serán discutidas, así como la utilidad de cada uno de los procedimientos.

La evaluación transcultural es actualmente uno de los temas más relevantes en la investigación psicológica y educativa. Muestra de ello son los programas internacionales diseñados para comparar personas de diferentes países, como por ejemplo el Programa Internacional para la Evaluación de Competencias en Adultos (PIAAC- Organisation for Economic Co-operation and Development, 2004) o el Programa para la Evaluación Internacional de Estudiantes (PISA-Organisation for Economic Co-operation and Development, 2006).

Los resultados de estos proyectos son utilizados frecuentemente para ordenar países en función de variables educativas o psicológicas, pero a pesar de la importancia de las decisiones que se toman, las comparaciones a veces se realizan sin tener en cuenta diferencias culturales o diferencias en los constructos y en los instrumentos que se emplean para medirlos. Este hecho es especialmente relevante cuando las comparaciones se realizan entre personas que responden a diferentes versiones lingüísticas. En estos casos, tal y como la International Test Commission (ITC) señala, la consistencia de la medida entre lenguas debe ser establecida (Hambleton, Merenda, y Spielberger, 2005; International Test Commission, 2010).

Una posible estrategia para evaluar la equivalencia de los ítems es el análisis del Funcionamiento Diferencial de los Ítems (DIF), que ha mostrado su utilidad para evaluar la validez de instrumentos utilizados en comparaciones internacionales. Aparece DIF cuando personas con el mismo nivel de la característica medida, pero que pertenecen a diferentes grupos, tienen distinta probabilidad de dar una determinada respuesta (Millsap y Everson, 1993).

Tradicionalmente, el análisis del DIF ha estado focalizado en la evaluación de grupos que responden a una misma versión del instrumento, por ejemplo mujeres y hombres. Sin embargo, recientemente el contexto de las comparaciones lingüísticas ha centrado la atención, por el hecho de que resulta más difícil mantener los supuestos de igualdad entre los grupos (Sireci, 2005).

Existen una gran variedad de técnicas para evaluar el DIF en ítems dicotómicos y politómicos (Hidalgo y Gómez-Benito, 2010). En este estudio se aplicaran el estadístico Mantel-Haenszel (MH) mediante el programa DIFAS (Penfield, 2005) y la Regresión logística Ordinal (RLO, Miller y Spray, 1993).

El objetivo del estudio es mostrar como analizar el DIF en ítems politómicos mediante MH y RLO. Para ello, se analizaron mediante MH y RLO las respuestas de participantes de Estados Unidos y de España a siete escalas incluidas en el Cuestionario del Estudiante del estudio PISA.

MÉTODO

Participantes

Para realizar este estudio, se analizaron las respuestas de 17,405 participantes de España (8,704 mujeres y 8,701 hombres) y 4,902 participantes de Estados Unidos (2,422 mujeres y 2,480 hombres). La edad de los participantes de España fue de 16 años en todos los casos, mientras que los participantes de Estados Unidos tenían entre 15 y 16 años (media 15.5 y DT 0.5).

Instrumentos

El análisis se realizó utilizando siete escalas del Cuestionario del Estudiante del Estudio PISA. Todos los ítems tenían formato tipo Likert con cuatro alternativas de

respuesta y evaluaban actitudes hacia la ciencia. En el estudio PISA estas escalas se dividen en cuatro temas principales: Interés, apoyo, motivación para aprender y auto-cogniciones. En total se analizaron 38 ítems.

Procedimiento

El primer paso del estudio fue seleccionar las escalas a analizar. Esta selección se llevó a cabo teniendo en cuenta dos criterios: la unidimensionalidad de la escala y la utilización de las puntuaciones totales, por parte de los investigadores del PISA, para informar sobre las actitudes de los estudiantes hacia la ciencia. Los datos fueron extraídos de la página web de la OECD (http://www.oecd.org).

En primer lugar, se eliminaron aquellos participantes con respuestas incompletas en una o más preguntas de las escalas, lo que supuso una reducción de aproximadamente el 12% de los datos en cada una de las muestras. A continuación se extrajeron dos submuestras aleatorias de la muestra española con el objetivo de conseguir grupos con tamaños comparables. También estás submuestras permitieron realizar dos comparaciones independientes entre los grupos que posteriormente fueron comparadas.

Análisis

Los análisis del DIF se aplicaron utilizando el «país» como variable de comparación y aplicando el esquema de *Funcionamiento Diferencial por Pasos* (DSF) desarrollado por Penfield para evaluar ítems politómicos (Penfield, 2010; Penfield, Gattamorta, & Childs, 2009). Se utilizó el programa DIFAS 4.0 (Penfield, 2007), mediante el cual se analizó primero el DIF a nivel del ítem, evaluando posteriormente el DSF solo en aquellos ítems detectados en la fase anterior. Para determinar los ítems con DIF se siguieron los criterios de Penfield y Algina (2003), y en el caso del análisis del DSF se utilizaron los parámetros determinados por Penfield, Gattamorta y Childs (2009) sobre el tamaño del efecto cuando se usan categorías acumulativas. Sólo los ítems detectados con tamaño del efecto medio o alto en ambas comparaciones fueron considerados como ítems con DIF.

Por otro lado, los análisis de la Regresión Logística Ordinal (RLO) fueron aplicados mediante el Statistical Package for Social Sciences (SPSS v.16) y siguiendo las indicaciones elaboradas por Zumbo (1999). El procedimiento fue el mismo, se analizó el DIF y el tamaño del efecto que fue clasificado utilizando los criterios de Penfield (2009) y el índice Delta de la Educational Testing Service (ETS).

RESULTADOS

Debido a la amplitud de los resultados obtenidos nos centraremos en los ítems de una de las escalas «valor general de la ciencia» que incluye cinco ítems. Se

muestran a continuación algunos análisis descriptivos y los resultados de los análisis del DIF. En primer lugar se muestra los resultados obtenidos para el MH en ambas réplicas y a continuación los resultados de la RLO. Por último, se incluye un ejemplo de los ítems detectados con DIF incluyendo las dos versiones y señalando aspectos que podrían estar provocando el DIF.

Estadísticos descriptivos

La tabla 1 muestra la media, la desviación típica (DT) y el índice de discriminación (ID) de los cinco ítems para los participantes de ambos grupos (España y Estado Unidos). También se incluyen los valores del coeficiente alfa de la escala para cada uno de los grupos.

Tabla 1. Estadísticos descriptivos de la escala «valor general de la ciencia»

Escala	Item	País					
		España			USA		
		Media	DT	ID	Media	DT	ID
Valor general de la ciencia (α España =.73; α EEUU=.82)	1	1.41	.58	.50	1.68	.65	.65
	2	1.56	.60	.48	1.61	.62	.62
	3	1.90	.74	.44	1.85	.68	.67
	4	1.82	.67	.48	1.76	.67	.66
	5	1.80	.69	.53	2.04	.76	.74

Como muestra la tabla 1, tanto los valores de la media como de la desviación típica fueron similares en ambos grupos, encontrándose las mayores diferencias en el ítem 5. En comparación con otras escalas, la escala «Valor general de la ciencia» obtuvo los valores más bajo en los índices de discriminación y el coeficiente alfa en ambos grupos.

Resultados del DIF: MH

La tabla 2 muestra los resultados de los análisis del MH. En esta tabla se especifica para cada una de las réplicas si el ítem ha sido detectado con DIF (S= si; N=no) y el tamaño del efecto obtenido (S=small, M=medium, y L=Large). El asterisco señala aquellos ítems detectados en ambas muestras con DIF medio o alto y por tanto considerados «ítems con DIF».

Como muestra la tabla 2, los resultados fueron bastante consistentes a través de las muestras. En esta escala, todos los ítems fueron detectados con DIF, lo que muestra la necesidad de aplicar otros métodos que nos permitan incrementar la seguridad en los resultados.

Table 2. Resumen de los resultados del MH

Escala	Item	Réplica			
		1		2	
		DIF	Efecto	DIF	Efecto
Valor general de la ciencia	1*	Y	L	Y	L
	2*	Y	L	Y	M
	3*	Y	L	Y	L
	4*	Y	L	Y	L
	5*	Y	L	Y	L

Resultados del DIF: RLO

En este caso los análisis del DIF siguieron la misma lógica que en el caso del MH. La única diferencia es que la clasificación del tamaño del efecto se realizó considerando dos criterios diferentes, el aplicado en el caso del MH lo que permite la comparabilidad entre los resultados de ambos procedimientos y el criterio de la ETS, el índice Delta. La tabla 3 muestra los resultados siguiendo el mismo esquema que la tabla 2.

Tabla 3. Resumen de los resultados de la RLO

Escala	Item	Réplica					
		1			2		
		DIF	Criterio 1.	Criterio ETS.	DIF	Criterio 1.	Criterio ETS.
Valor general de la ciencia	1*	Y	L	L	Y	L	L
	2	Y	L	L	N		
	3*	Y	M	M	Y	M	M
	4	Y	M	M	Y	M	S
	5*	Y	L	L	Y	L	L

Como muestra la tabla 4, en este caso sólo tres ítems fueron detectados en todas las condiciones con DIF medio o alto.

Por lo tanto, teniendo en cuenta los resultados de ambos procedimientos, en esta escala los ítems considerados como «ítems con DIF» serían el ítem 1, el ítem 3 y el ítem 5. En el total de las escalas, 16 ítems de 38 fueron detectados usando estos criterios.

Ejemplo ítems

Para ilustrar esta situación que recogen los resultados, la tabla 4 presenta las versiones de los ítems 1, 3 y 5 que respondieron los estudiantes españoles y los de Estados Unidos.

Tabla 4. Versiones de los ítems detectados con DIF

Item	Versión para España	Versión para EEUU
Item 1	Los avances en ciencia y tecnología suelen mejorar las condiciones de vida de las personas.	Advances in science and technology usually improve people's living conditions.
Item 3	Los avances en ciencia y tecnología ayudan a mejorar la economía.	Advances in science and technology usually help improve the economy.
Item 5	Los avances en ciencia y tecnología suelen proporcionar beneficios sociales.	Advances in science and technology usually bring social benefits.

Como muestra la tabla 4, en los tres ítems detectados aparece una expresión común «Los avances en ciencia y tecnología». Esta expresión podría estar influenciando la interpretación realizada por los grupos de forma que apareciera DIF. Un grupo de expertos podría ser útil para evaluar diferencias en las versiones que pudieran estar provocando interpretaciones diferentes.

DISCUSIÓN

El objetivo de este estudio era ilustrar la evaluación del DIF en ítems politómicos actitudinales mediante MH y RLO. Los resultados han mostrado la necesidad de implementar diferentes métodos de forma que se pueda tener seguridad en las conclusiones establecidas ya que, a pesar de ser ésta la escala con mayor número de ítems con DIF, los resultados proporcionados por el MH detectaron cinco ítems (todos) con DIF, dos de los cuáles se eliminaron de este grupo al analizar los datos de la RLO.

Es importante también destacar la importancia de aplicar análisis del DIF en estudios transculturales. La existencia de ítems con DIF supone una debilidad a la hora de comparar los resultados de los participantes de ambos grupos, que puede provocar que por ejemplo ordenemos los países basándonos en criterios inadecuados.

Por último, comentar que este estudio forma parte de un proyecto cuyo objetivo es la búsqueda de las causas del DIF. Los resultados del estudio se utilizarán para guiar el diseño y la realización de entrevistas cognitivas que estarán focalizadas en localizar patrones de interpretación diferentes entre los grupos.

REFERENCES

Hambleton, R. K., Merenda, P., y Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Lawrence S. Erlbaum.

Hidalgo, M. D., y Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (3rd edition)*. USA: Elsevier - Science & Technology.

International Test Commission (2010). *Guidelines for translating and adapting tests*. Downloaded from the world wide web at http://www.intestcom.org on October 4, 2010.

Miller, T.R. y Spray, J.A. (1993). Logistic Discriminant Function Analysis for DIF Identification of polytomously scored items. *Journal of Educational Measurement 30 (2)*, 107-122.

Millsap, R. E. y Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement 17***,** 297-334.

Organisation for Economic Co-operation and Development. (2004). Programme for the International Assessment of Adult Competencies (PIAAC). Policy Objectives, Strategic Options and Cost Implications. Stockholm: Author.

Organisation for Economic Co-operation and Development. (2006). Literacy skills for the world of tomorrow—further results from PISA 2003. Paris: Author.

Penfield, R. D. y Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40,* 353-370.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150-151.

Penfield, R. D. (2007). *DIFAS 4.0 user's manual*. Downloaded from the world wide web on July 14, 2010 from

http://www.education.miami.edu/facultysites/penfield/index.html.

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement*, *47*, 129-149.

Penfield, R. D., Gattamorta, K. y Childs, R. A. (2009), An NCME Instructional Module on Using Differential Step Functioning to Refine the Analysis of DIF in Polytomous Items. *Educational Measurement: Issues and Practice, 28*, 38–49.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Hillsdale, NJ: Lawrence Erlbaum.

SPSS, Inc. 2007. SPSS-16 User's guide. Chicago, USA.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

COMPLEMENTARIEDAD METODOLÓGICA

Coordinadora: M. Teresa Anguera
Universidad de Barcelona

La historia del desarrollo científico en diversas áreas del conocimiento revela diversas estrategias en el estudio del comportamiento, y destaca principalmente la confrontación entre las estrategias cualitativa y cuantitativa. Durante décadas se produjo un movimiento pendular que otorgaba relevancia a una u otra, y criticaba duramente a la opción alternativa. En los últimos años se ha ido configurando cada vez con mayor fuerza la metodología mixta se caracteriza, que no se limita a la simple recogida de datos de diferente naturaleza sobre el mismo comportamiento o episodio, sino que implica, por una parte, mezclar la lógica inductiva y la deductiva, y, por otra, mantener este carácter mixto a lo largo de todo el proceso, es decir, en el planteamiento del problema, recogida de datos, análisis de datos, e interpretación de resultados, y se debe manifestar en el informe científico que se realice. Precisamente porque la metodología mixta tiene una presencia integral en el desarrollo metodológico 'completo', resulta esencial distinguir los métodos, diseños y medidas, como tres componentes fundamentales dispuestos de forma concatenada en planos jerárquicos diferenciados; este aspecto fundamental es el que otorga su carácter singular y diferenciador, por cuanto no se trata de una simple yuxtaposición de etapas sino de un modo o manera particular de proceder en el abordaje del conocimiento. Es por ello, que la complementariedad metodológica se convierte en la «clave» o «piedra angular» de esta perspectiva y justifica que algunos autores la consideren como la «tercera vía» o alternativa metodológica. De cualquier modo, queremos enfatizar que ésta, ha de entenderse como un planteamiento integrador y, su fortaleza la encuentra en el uso conjunto de ambas metodologías y no como una simple posición ecléctica ante un tercer paradigma combinatorio de las excelencias de datos cuantitativos e informaciones cualitativas. En este Symposium se presentan cinco trabajos que materializan la complementariedad metodológica, tanto en el planteamiento del diseño, como en los instrumentos de medida utilizados, como en los datos obtenidos, como en las técnicas de análisis utilizadas.

PALABRAS CLAVE: Diseños mixtos, Diseños de investigación híbridos, Diseños integrados de dominancia, Unidades multimodales.

REVISIÓN DE LAS PROPUESTAS DE LOS DISEÑOS DE INVESTIGACIÓN HIBRIDOS EN CIENCIAS SOCIALES Y DEL COMPORTAMIENTO

Olatz López-Fernández y Mª Teresa Anguera-Argilaga
Universidad de Barcelona

Resumen

Recientemente se ha empezado a consolidar el *mixed methods approach* como una opción integradora de la investigación cuantitativa y cualitativa en el campo de las ciencias sociales y del comportamiento. Esta corriente metodológica ha proporcionado un nuevo marco en el que ubicar investigaciones que contemplan la complementariedad de los dos enfoques clásicos a lo largo del proceso de investigación, dando respuesta a una demanda existente desde hace al menos dos décadas. La institucionalización de los *mixed methods* a partir de 1990 como tercera orientación metodológica en las ciencias citadas, ha dado lugar a un nuevo paradigma metodológico que desde 2003 ha emergido con entidad propia, con una serie de elementos metodológicos únicos que tienen una terminología específica. En concreto, uno de los aspectos más relevantes son los diseños de *mixed methods research* (también denominados diseños de investigación híbrida), en que hasta el momento se contemplan diversas clasificaciones que han ido madurando con el tiempo y dando respuesta a las posibles situaciones en que pueden encontrarse las investigaciones que impliquen la complementariedad de métodos cuantitativos y cualitativos. En el presente trabajo se van a desarrollar con orden cronológico las distintas propuestas actualmente utilizadas, así como la forma de representación de dichos diseños con su terminología para discutir sus principales bondades e inconvenientes metodológicos.

La utilización de los *mixed methods research,* o metodología híbrida, hace referencia a la combinación de métodos cuantitativos y cualitativos en una misma investigación, a modo de complementariedad metodológica ampliamente utilizada en las ciencias sociales y del comportamiento. Los métodos híbridos parecen haber existido incluso antes de que Tashakkori y Teddlie (1998; 2003; 2010) publicaran los *handbook* que han marcado el inicio de esta tercera aproximación metodológica, por detrás de la cuantitativa, en primer lugar, y la cualitativa, en segundo lugar. La aportación que realizan los investigadores del ámbito de la metodología, que a principios de la última década se dedican a desarrollar esta aproximación, básicamente consiste en marcar las pautas generales que permiten dar nombre a este tipo de investigaciones, así como llegar a una serie de acuerdos internacionales respec-

to a qué son los *mixed methods* studies, cómo deber ser desarrollados atendiendo a la combinación de ambas aproximaciones, con qué propósito de complementariedad metodológica, con qué tipos de diseños *mixed methods* (o diseños híbridos), con qué tipo de técnicas de muestreo híbridas, con qué tipo de instrumentos y procedimientos propios de ambos métodos, con qué tipo de técnicas de análisis de datos cuantitativos y cualitativos, así como elementos relacionados en cómo mostrar los resultados híbridos para interpretarlos y asignarles su aportación en relación a la validez de este tipo de estudios.

En este momento, de más de una década de desarrollo de los métodos híbridos, parece necesario darlos a conocer y clasificarlos cronológicamente para entender de donde parte este esfuerzo en consolidar la complementariedad metodológica a nivel internacional. Por ello, este trabajo muestra la línea que se desarrolla actualmente en relación a conocer, en detalle, algunas de las principales características de los elementos metodológicos de la investigación con metodología híbrido para ser aplicadas especialmente en las ciencias sociales y del comportamiento.

Nuestro propósito es desarrollar, con orden cronológico, algunas de las principales propuestas actualmente utilizadas en metodología híbrida, así como la forma de representación de dichos diseños con su terminología internacional. Nuestra finalidad es realizar esta revisión para que investigadores de las ciencias sociales y del comportamiento, entre otras disciplinas, puedan conocer las posibilidades de esta aproximación de la complementariedad metodológica presente, además de poder utilizarla en sus respectivas investigaciones con los códigos internacionales para facilitar la difusión de sus trabajos.

En síntesis, se resaltaran los diseños de metodología híbrida más relevantes y se valorarán los aspectos más favorables y los más críticos de las propuestas existentes hasta el momento.

Se sigue la estructura siguiente: se presentan los «Diseños híbridos clásicos» (1991-2004), a continuación los «Diseños híbridos bidimensionales» (2006), los «Diseños híbridos con los cuatro tipos principales» (2007), los «Diseños híbridos según tipología tridimensional» (2009) y finalmente se procede a la «Discusión y conclusiones» de esta revisión cronológica de los diseños de investigación híbridos.

DISEÑOS HÍBRIDOS CLÁSICOS

En el trabajo de Molina-Azorín y López-Fernández (2009) se presentó la clasificación clásica de los diseño híbridos establecida entre la primera publicación de Tashakkori y Teddlie (1998), en que por primeva vez se nombra a la «mixed methodology» como tercera aproximación metodológica, hasta la publicación del *handbook of mixed methods in social and behavioral research* de los mismos autores (Tashakkori y Teddlie, 2003), que establece los fundamentos de la metodología híbrida a nivel internacional. En este mismo trabajo, los autores citaban como elementos claves el *propósito* del diseño híbrido (según Greene, Caracelli & Gra-

ham, 1989), que podía ser la triangulación, el desarrollo, la complementariedad o la expansión, así como los dos ejes que iban a determinar el tipo de diseño híbrido clásico:

- Prioridad (Creswell, 2003; Morgan, 1998; Morse, 1991): diferente o equivalente
- Implementación (Creswell, 2003; Morgan, 1998; Morse, 1991): secuencial o simultáneo

Pues de ambos ejes surgían los nueve tipos de diseño híbrido más utilizados hasta la actualidad (según Creswell, 2003; Morgan, 1998; Morse, 1991; Johnson y Onwuegbuzie, 2004) (figura 1).

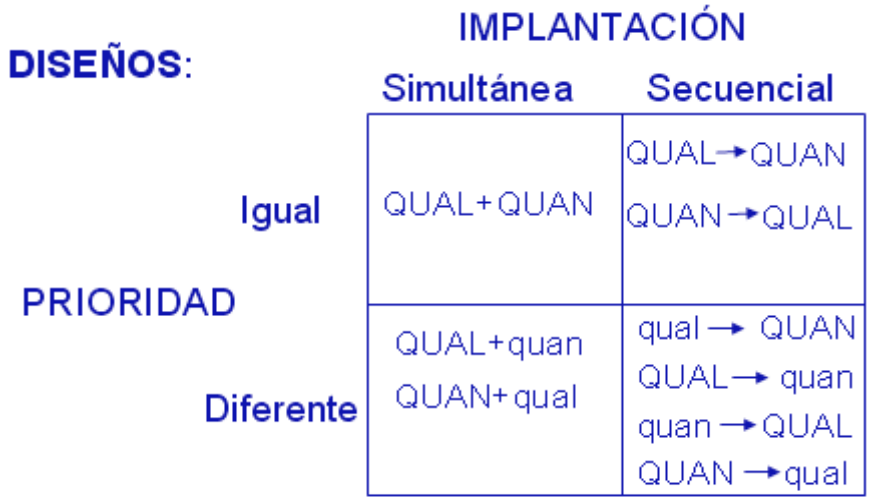


Figura 1. Nueve tipos de diseños híbridos clásicos

DISEÑOS HÍBRIDOS BIDIMENSIONALES

No obstante, nuevos avances se han ido realizando en relación a aspectos relacionados con los diseños híbridos, como la clasificación de Collins, Onwuegbuzie y Jiao (2007), que extraen ocho diseños híbridos en relación a los tipos de muestreo (figura 2).

En este caso, se basan en la notación de Morse (1991) y en el eje de implementación (Creswell, 2003; Morgan, 1998) para determinar en función de ambos a cada tipo de diseño en función de si el muestreo era idéntico, paralelo, anidado o multinivel. Por ello lo denominan «modelo de muestreo híbrido bidimensional».

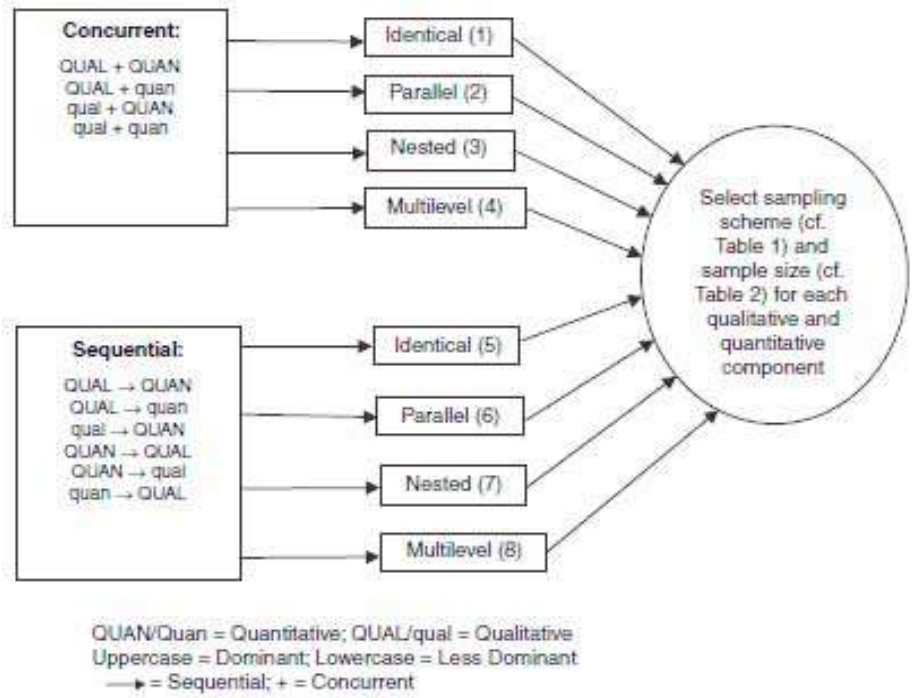


Figure 2 Two-dimensional mixed-methods sampling model providing a typology of mixed-methods sampling designs

Figura 2. Ocho tipos de diseños híbridos bidimensionales

DISEÑOS HÍBRIDOS CON LOS 4 PRINCIPALES TIPOS

Posteriormente, Creswell y Plano-Clark (2007) establecen otra clasificación con cuatro diseños híbridos principales (figura 3).

Esta tercera clasificación ha sido bastante utilizada (después de la primera clasificación clásica), pues describe los diseños híbridos de triangulación, explicativo, exploratorio e incrustado, tratando de relacionar el propósito de diseño híbrido con la forma de mezclar ambas aproximaciones, la cuantitativa y cualitativa, en una secuencia lógica.

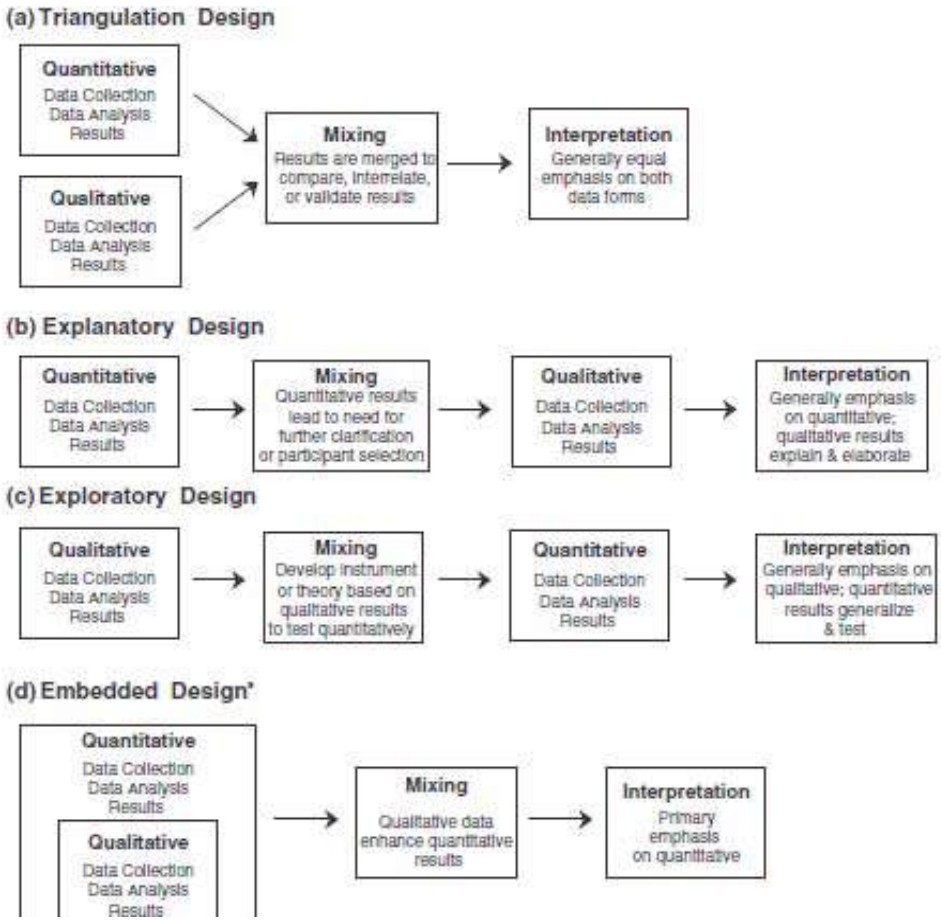


Figura 3. Cuatro tipos de diseños híbridos

DISEÑOS HÍBRIDOS SEGÚN TIPOLOGÍA TRIDIMENSIONAL

Leech y Onwuegbuzie (2009) desarrollaron una propuesta (figura 4) que también está siendo utilizado en el presente, al establecer ocho diseños híbridos en relación a tres ejes: la dimensión de mezcla (total o parcial), la dimensión del tiempo (concurrente o secuencial) y la dimensión del énfasis (igual o diferente estatus).

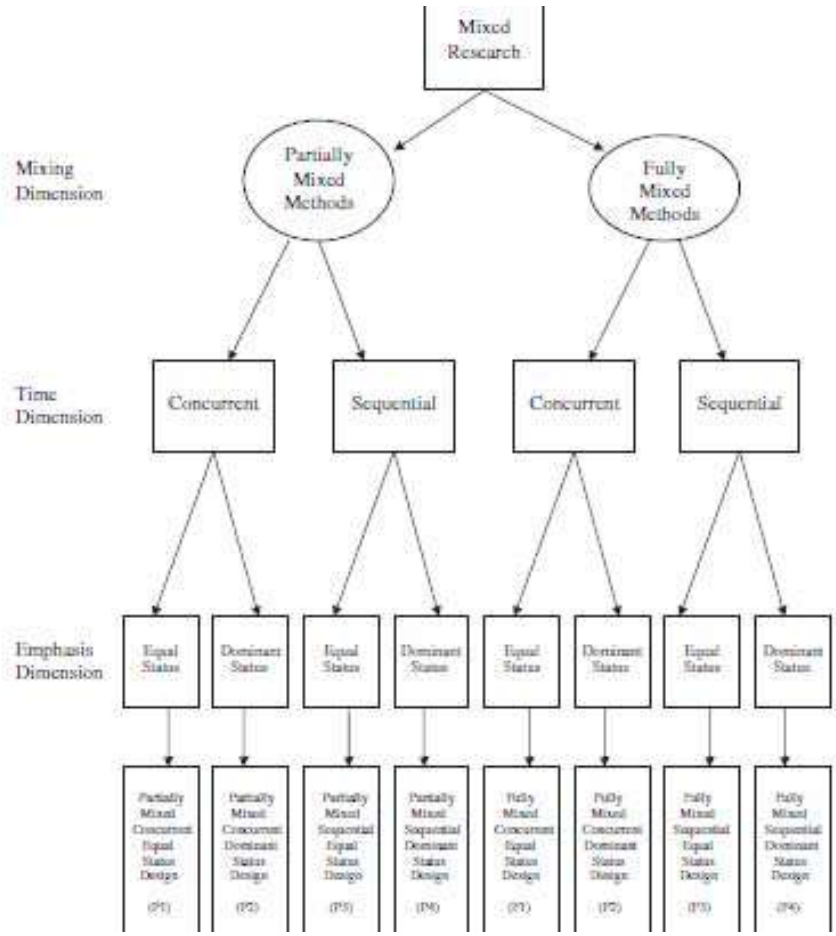


Figura 4. Ocho tipos de diseños híbridos tridimensionales

DISCUSIÓN Y CONCLUSIONES

En primer lugar, se destacar el progresivo desarrollo de propuestas relacionadas con los avances de los diseños híbridos, todas ellas utilizadas en los trabajos con metodología híbrida, cuyo uso y publicación presenta un crecimiento relevante entre el período 2003-2010. Parece lógico pensar que los diseños han de evolu-

cionar hasta tratar de dar respuesta a todas las posibilidades metodológicas posibles de la combinación de las aproximaciones cuantitativa y cualitativa, y en este sentido, el nuevo *handbook* de Tashakkori y Teddlie (2010) aborda en detalle este aspecto, así como publicaciones recientes en relación a este tipo de diseños de investigación (Morse y Niehaus, 2009). No obstante, excesiva complejidad puede dispersar un lenguaje común y sencillo como el que se había establecido con la clasificación clásica.

En segundo lugar, se han presentado las principales clasificación de diseños híbridos publicadas y utilizadas hasta el momento, aunque no son exclusivas, dado que los autores tienden a hacer matices e incluir algún aspecto propio en sus investigaciones, en parte para aclarar aspectos que pueden no quedar del todo visibles al nombrar el diseño híbrido con la denominación y notación establecidas.

En tercer lugar, siguen predominando los diseños clásicos en revisiones y estudios de prevalencia La notación de Morse (1991) sigue predominando en todas las clasificaciones. Las nuevas clasificaciones tienden a plantear nuevas posibilidades (i.e: *nested design*), así como tienden a tratar elementos más complejos del diseño híbrido.

En cuarto lugar, hace falta más investigación en diseños híbridos para poder, no solo dar respuesta a los aspectos metodológicos de la complementariedad, sino también para generar un corpus metodológico que permita que investigadores de las ciencias sociales y del comportamiento puedan dar difundir sus resultados y, a su vez, establecer las bondades (e inconvenientes) que puede proporcionar la elección de un diseño híbrido determinado. En nuestro caso, centramos la atención en la metodología híbrida en Psicología.

En esta línea, en especial desde el ámbito de las ciencias del comportamiento seguimos trabajando para poder dar algunas de estas respuestas a nuestros colegas, así como a investigadores de otros ámbitos como son las ciencias de la educación y de la salud, donde la metodología híbrida parece estar dando respuestas a nivel de la investigación internacional.

REFERENCIAS

Creswell, J. (2003). *Research design. Qualitative, quantitative and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. y Plano-Clark, V. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Creswell, J., Plano Clark, V., Gutmann, M. y Hanson, W. (2003). Advanced mixed methods research designs. En A. Tashakkori y C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 209-240). Thousand Oaks, CA: Sage.

Collins, K. T., Onwuegbuzie, A. J., y Jiao, Q. G. (2007). A Mixed Methods Investigation of Mixed Methods Sampling Designs in Social and Health Science Research. *Journal Of Mixed Methods Research, 1*(3), 267-294.

Greene, J., Caracelli, V. y Graham, W. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11,* 255-274.

Jick, T. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly, 24,* 602-611.

Johnson, B. y Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Leech, N. L., y Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality & Quantity: International Journal Of Methodology, 43*(2), 265-275.

Molina-Azorín, J.F. y López-Fernández, O. (2009). Revisión y comparación de la aplicación de la metodología híbrida en tres revistas de referencia en ciencias del comportamiento. En el *Libro de Actas del XI Congreso de Metodología de las Ciencias Sociales y de la Salud*, en la Universidad de Málaga, Facultad de Psicología, Departamento de Metodología de las Ciencias del Comportamiento, 15-18 de Septiembre de 2009.

Morgan, D. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative Health Research, 8,* 362-376.

Morse, J. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research, 40,* 120-123.

Morse, J.M. y Niehaus, L. (2009). *Mixed methods design. Principles and procedures*. Walnut Creek, CA: Left Coast Press.

Tashakkori A. y Teddlie C. (1998) *Mixed methodology. Combining qualitative and quantitative approache*s. Thousand Oaks, CA: Sage.

Tashakkori, A. y Teddlie, C. (Eds.) (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage.

Teddlie, C. y Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13,* 12-28.

APLICACIÓN DE LA METODOLOGÍA HÍBRIDA EN PSICOTHEMA (2003-2010): PROPÓSITOS, DISEÑOS Y RECOMENDACIONES

Olatz López-Fernández[1] José Francisco Molina-Azorín[2]
[1] Universidad de Barcelona
[2] Universidad de Alicante
Correo electrónico: olatzlopez@ub.edu

Resumen

La metodología híbrida (MH) (*mixed methods*) consiste en el uso combinado de métodos cuantitativos y cualitativos en un mismo trabajo de investigación. Esta tradición metodológica está adquiriendo importancia en el área de las ciencias sociales y del comportamiento, y se le está llegando a considerar como una tercera aproximación metodológica, por constituir una alternativa a la metodología cuantitativa (QUAN) y la cualitativa (QUAL). A diferencia de estas tradiciones clásicas, la MH se basa en un paradigma pragmático con preguntas de investigación QUAL y QUAN, que requieren de métodos mixtos, con diseños híbridos propios y un potente nivel de inferencia. El objetivo principal de este trabajo es examinar los artículos publicados en los últimos ocho años (2003-2010) en Psicothema y clasificarlos en teóricos o empíricos, y dentro de estos últimos, en cuantitativos, cualitativos e híbridos. Además, para cada artículo híbrido identificado se analizan sus principales características en cuanto al propósito metodológico y al diseño híbrido de investigación empleado, lo que permitirá observar la evolución de las metodologías implementadas en los trabajos publicados en éstos últimos años en que la MH aparece como una opción válida a la complementariedad metodológica del área de ciencias del comportamiento. También se señalarán recomendaciones para aplicar los mixed methods en esta área de conocimiento para aprovechar todo su potencial en futuras investigaciones.

La utilización de métodos híbridos cuantitativos y cualitativos en una misma investigación está adquiriendo una creciente importancia (Creswell, 2003; Tashakkori y Teddlie, 2003) en las ciencias sociales y del comportamiento. La utilización de métodos híbridos puede jugar un rol importante en la investigación, al enriquecer y mejorar la comprensión del fenómeno estudiado. Creswell y Plano Clark (2007) indican que la aplicación y uso de la metodología híbrida se ha revisado en pocas disciplinas, por lo que un tema de interés es cómo está siendo aplicada en otras áreas. El objetivo principal de este trabajo es examinar los artículos publicados en los últimos ocho años (2003-2010) en Psicothema y clasificarlos en teóricos

o empíricos, y dentro de estos últimos, en cuantitativos, cualitativos e híbridos. Además, para cada artículo híbrido identificado se analizan sus principales características en cuanto al propósito metodológico y al diseño híbrido de investigación empleado.

MÉTODOS DE INVESTIGACIÓN HÍBRIDOS

Las disciplinas en que está aumentando la literatura relativa a la metodología híbrida y su estudio son la evaluación, educación, sociología, ciencias de la salud y psicología (Tashakkori y Teddlie, 2003). De hecho, sobre esta aproximación metodológica se están publicando revistas específicas dedicadas a esta aproximación, como *Journal of Mixed Methods Research* o *International Journal of Multiple Research Approaches,* la primera de ellas con factor de impacto en el ISI JCR desde 2010.

Los dos factores que ayudan a determinar los diseños híbridos más comúnmente utilizados en metodología híbrida son (Creswell, 2003; Morgan, 1998; Morse, 1991):

– Prioridad/importancia. Consiste en dar la misma prioridad o importancia a las partes cuantitativa y cualitativa, o bien otorgar mayor prioridad o relevancia a una de ellas.

– Implantación de la recogida de datos. Se refiere a la secuencia con que el investigador recoge los datos cuantitativos y cualitativos, que puede ser de dos maneras: al mismo tiempo (diseño simultáneo, concurrente o paralelo) o en diferentes fases a lo largo del tiempo (diseño secuencial o en dos o más fases).

Para la representación de estos diseños se suele utilizar la notación de Morse (1991). En su sistema utiliza la abreviatura «quan» para representar la parte cuantitativa y «qual» para la cualitativa. Además, cuando hay un método dominante o más importante, éste se representa con letras mayúsculas (QUAN, QUAL) mientras que el método menos importante aparece con minúsculas (quan, qual). Respecto a la implantación de recogida de datos, el símbolo «+»indica un diseño simultáneo, mientras que la flecha «→» se refiere a un diseño secuencial. Así, podemos establecer cuatro bloques que dan lugar a nueve diseños híbridos (Johnson y Onwuegbuzie, 2004) (figura 1).

En cuanto a los propósitos de los diseños híbridos (Creswell, 2003; Greene, Caracelli y Graham, 1989; Morgan, 1998), los más extendidos son la triangulación y la complementariedad. Con la *triangulación* (Jick, 1979) se persigue una convergencia de los resultados a partir de ambas aproximaciones cuantitativa y cualitativa, para conseguir mayor fiabilidad, es decir, se busca la corroboración de resultados desde diferentes métodos. En cambio, la *complementariedad* busca que una de las aproximaciones (la cuantitativa o la cualitativa) complemente a la otra, esto es, clarificar, mejorar, ampliar o ilustrar a través de uno de los métodos los resultados

obtenidos en el otro método. Otro tercer propósito es el de *desarrollo*, en que uno de los métodos (el menos importante) ayuda en algún aspecto a mejorar la ejecución posterior del otro método (el método principal). Un cuarto propósito es el de *expansión*, que trata de buscar un análisis y comprensión de diferentes facetas de un fenómeno, obteniendo así una comprensión más rica y profunda del fenómeno.

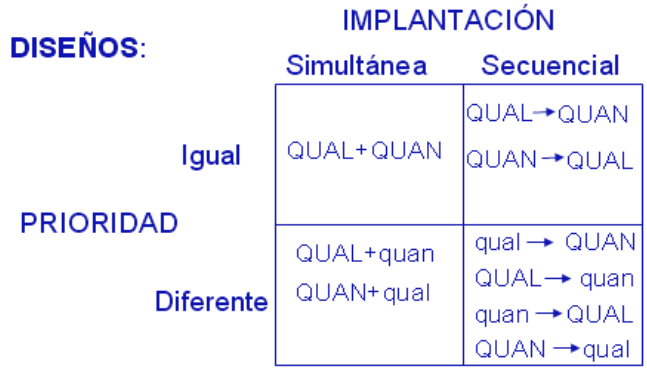


Figura 1. Tipos de diseños híbridos

METODOLOGÍA

Con el objetivo de identificar y analizar los diseños híbridos utilizados y los propósitos perseguidos en el campo de las ciencias del comportamiento, hemos llevado a cabo una revisión de los artículos publicados en tres revistas de referencia en diversas temáticas de este campo, en concreto *Psicothema*.

Para identificar los estudios híbridos publicados en estas revistas, todos los artículos publicados desde 2003 a 2010 fueron revisados, de modo que fueron analizados 8 Volúmenes, 32 números y 959 artículos. El inicio en 2003 se debe a que es el año en que se publicó el *Handbook of Mixed Methods in Social and Behavioral Research* de Tashakkori y Teddlie, que ha proporcionado visibilidad y credibilidad a este enfoque. Esta estrategia de búsqueda nos permitió, además de identificar los estudios híbridos, clasificar todos los trabajos en dos grupos, no empíricos y empíricos, y desdoblar este grupo de artículos empíricos en tres tipos: cuantitativos, cualitativos e híbridos. Una vez realizada esta clasificación, se realizó un análisis de contenidos de los artículos híbridos identificados, determinando para cada uno de ellos el tipo de diseño utilizado en función de las características de prioridad e implantación previamente señaladas así como el propósito híbrido principal del trabajo.

Cabe indicar que determinar si un estudio es híbrido implica la lectura completa del artículo, especialmente del apartado del método, dado que la información referente al diseño apenas aparece en los resúmenes y por el momento parece que se observa poca bibliografía en relación a la metodología híbrida a pesar de que se utiliza cada vez con mayor frecuencia.

RESULTADOS

La tabla 1 muestra el tipo de artículos identificados en Psicothema. En esta tabla se observa, en primer lugar, el predominio de artículos empíricos (88,4%) frente a los no empíricos (11,6%). Además, esta revista tiene una clara orientación a publicar trabajos empíricos de tipo cuantitativo (823, 85,8%), mientras que los estudios cualitativos e híbridos son minoritarios (1,5% y 1,1% respectivamente).

Con relación a la publicación de estudios híbridos, también se observa en la tabla 1 que no existe un clara tendencia creciente hacia la aplicación de esta metodología. De hecho en el primer año analizado (2003) es cuando se ha publicado un mayor número de estudios híbridos, mientras que en el último año examinado (2010) no se ha identificado ningún artículo híbrido.

Tabla 1. Tipos de artículos publicados en Psicothema (2003-2010)

Año	Número total de artículos	Número de artículos no empíricos	Artículos empíricos			
			Número total de artículos empíricos	Número de artículos cuantitativos	Número de artículos cualitativos	Número de artículos híbridos
2003	101	12	89	82	3	4
2004	101	17	84	80	4	0
2005	106	14	92	87	2	3
2006	152	15	137	134	1	2
2007	102	13	89	89	0	0
2008	146	10	136	136	0	0
2009	97	7	90	88	0	2
2010	154	23	131	127	4	0
TOTAL	959	111	848	823	14	11

Con relación a los estudios híbridos, en la tabla 2 se sintetizan tanto las características de los diseños híbridos como los propósitos metodológicos de este tipo de artículos que integran métodos cuantitativos y cualitativos.

Como se observa en la tabla 2, con relación al tipo de prioridad, predominan los artículos híbridos donde la prioridad es equivalente, es decir, donde las partes cuantitativa y cualitativa tienen la misma relevancia. En cuanto al tipo de implantación, también hay una clara orientación hacia una implantación de tipo secuencial, donde la primera parte es cualitativa y la segunda cuantitativa.

Combinando las dos anteriores características (diseño e implantación) se ha encontrado que el mayor número de estudios híbridos son de tipo equivalente / secuencial. En concreto, se han identificado seis estudios con diseño QUAL→QUAN. A continuación, se han identificado 3 estudios de prioridad diferente e implantación secuencial, concretamente 3 estudios qual→QUAN. Por último, 2 estudios presentaban un diseño prioridad equivalente e implantación simultánea (QUAL+QUAN).

Tabla 2. Características de los artículos híbridos identificados en Psicothema (2003-2010)

	Número de artículos (%)
Prioridad	
Equivalente	8 (72.7%)
Diferente	3 (27.3%)
Implantación	
Simultánea	2 (18.2%)
Secuencial	9 (81.8%)
Diseño	
Equivalente / simultáneo	2 (18.2%)
Equivalente / secuencial	6 (54.5%)
Diferente / simultáneo	0 (0%)
Diferente / secuencial	3 (27.3%)
Propósitos	
Triangulación	0 (0%)
Complementariedad	0 (0%)
Desarrollo	9 (81.8%)
Expansión	2 (18.2%)
TOTAL	11 (100%)

Finalmente, analizando los propósitos perseguidos, 9 estudios híbridos perseguían el propósito de desarrollo y 2 artículos es de expansión. No se ha encontrado ningún estudio híbrido con los propósitos de triangulación y complementariedad.

RECOMENDACIONES Y CONCLUSIONES

Como conclusiones, podemos indicar varias ideas principales. En primer lugar, hemos de destacar la poca utilización de la metodología híbrida en Psicothema en el período 2003-2010. Sólo 11 trabajos de un total de 959 publicados (1.1%) han utilizado esta metodología. En segundo lugar, ninguno de estos trabajos utiliza la expresión «métodos híbridos o mixtos» a lo largo del trabajo. Y, además, revisando la lista de referencias bibliográficas de estos trabajos, ningún artículo híbrido identificado ha citado referencias clásicas vinculadas a la metodología híbrida. Es por ello que podría pensarse que esta metodología, que en otros campos y países tiene una identidad propia reconocida, no es completamente conocida por los investigadores en nuestro país, y por tanto, quizás se está dejando de aprovechar todo el potencial que presenta para analizar problemas complejos de investigación.

La realización de trabajos híbridos en el campo de las ciencias del comportamiento en nuestro país puede aprender determinados aspectos de los trabajos ya realizados en este y otros campos en otros países. En este sentido, Creswell, Plano

Clark, Gutmann y Hanson (2003) señalan que un reto importante para los trabajos híbridos es la clarificación explícita de varios aspectos relevantes. En primer lugar, los investigadores deberían identificar claramente cuáles son las razones o propósitos principales de utilizar en su trabajo un diseño híbrido utilizando datos cuantitativos y cualitativos, así como las principales ventajas y obstáculos de llevarlos a cabo. Además, también debería clarificarse los factores que hemos analizado hasta ahora para determinar los tipos de diseños. Así, con relación a la prioridad, los investigadores deberían indicar claramente las decisiones tomadas relativas a la importancia o atención prestada a la parte cuantitativa y cualitativa (igual o distinta importancia), lo cual podría reflejarse en la longitud y profundidad de los comentarios y discusiones realizados para cada uno de las aproximaciones. Por otra parte, con relación a la implantación de la recogida de datos, debería determinarse claramente si el diseño es secuencial o simultáneo. Por ejemplo, si el diseño es secuencial, las dos fases de recogida y análisis de datos podrían aparecer en el documento escrito de forma separada, llevando a cabo la integración de información en las secciones de discusión y/o conclusiones. Dada la complejidad de estos aspectos, el investigador puede utilizar figuras o modelos visuales para presentar su trabajo.

Además, los investigadores deben tener en cuenta que para un determinado estudio en principio puede establecer un diseño concreto, pero que nuevos componentes o ideas pueden surgir conforme el trabajo se va realizando, lo que puede conllevar una modificación de ese diseño previo. De esta forma, el investigador debe ser creativo y no debe limitarse a los diseños preexistentes, sino que debe incluso crear aquellos diseños que permitan contestar adecuadamente sus cuestiones de investigación. En esta línea, normalmente se señala que los diseños híbridos secuenciales tienen dos partes o fases, pero estos diseños pueden ser más complejos implicando tres o más fases (Johnson y Onwuegbuzie, 2004; Teddlie y Tashakkori, 2006).

Pensamos que como futuras líneas de investigación puede ser interesante ampliar el período analizado y realizar un análisis de otras revistas para tener una visión más amplia de la aplicación de la metodología híbrida en este ámbito. En nuestra opinión, este trabajo puede ayudar a difundir la utilización de la metodología híbrida en las ciencias del comportamiento. Los investigadores deben conocer la aceptación de esta metodología en sus respectivas áreas y la utilización de la misma por otros colegas.

REFERENCIAS

Creswell, J. (2003). *Research design. Qualitative, quantitative and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. y Plano-Clark, V. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Creswell, J., Plano Clark, V., Gutmann, M. y Hanson, W. (2003). Advanced mixed methods research designs. En A. Tashakkori y C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 209-240). Thousand Oaks, CA: Sage.

Greene, J., Caracelli, V. y Graham, W. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11,* 255-274.

Jick, T. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly, 24,* 602-611.

Johnson, B. y Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Morgan, D. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative Health Research, 8,* 362-376.

Morse, J. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research, 40,* 120-123.

Tashakkori, A. y Teddlie, C. (Eds.) (2003). *Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks, CA: Sage.

Teddlie, C. y Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13,* 12-28.

SESIONES PARALELAS

APLICACIONES PSICOMÉTRICAS

RELACIÓN DEL BURNOUT CON LOS ESTADOS DE ÁNIMO, LA ANSIEDAD Y LA AUTOCONFIANZA DE LOS DEPORTISTAS

Constantino Arce, Mar Graña, Cristina de Francisco e Iria Arce
Universidad de Santiago de Compostela
Correo electrónico: constantino.arce@usc.es

Resumen

El Athlete Burnout Questionnaire (ABQ) es un cuestionario elaborado por Raedeke y Smith (2001) para evaluar el burnout en deportistas, que fue adaptado al español por investigadores de la Universidad de Santiago de Compostela. El presente trabajo se realizó con el objetivo fundamental de estudiar el grado de relación o asociación existente entre el burnout, medido a través del ABQ, los estados de ánimo, la ansiedad y la autoconfianza de los deportistas. El ABQ mostró índices de consistencia interna aceptables en todos sus factores ($\alpha > .70$). Las puntuaciones en burnout, obtenidas con el ABQ, reflejaron relaciones directas con los estados de ánimo negativos y con la ansiedad somática pero no con la ansiedad cognitiva. También mostraron relaciones inversas con los estados de ánimo positivos y con la autoconfianza. No se encontraron diferencias de género en los niveles de burnout pero sí importantes diferencias entre grupos de edad, aumentando los niveles de burnout con la edad de los deportistas.

El Athlete Burnout Questionnaire (ABQ) es un cuestionario elaborado por Raedeke y Smith (2001) para evaluar el burnout en deportistas, que fue adaptado al español por investigadores de la Universidad de Santiago de Compostela, en colaboración con el primer autor, mediante análisis factoriales confirmatorios del modelo original con diversas muestras de deportistas (de Francisco, Arce, Andrade, Arce y Raedeke, 2009; Arce, de Francisco, Andrade, Arce y Raedeke, 2010; de Francisco, 2010). El presente trabajo se realizó con el objetivo fundamental de estudiar el grado de relación o asociación existente entre el burnout, medido a través del ABQ, los estados de ánimo, la ansiedad y la autoconfianza de los deportistas. Adicionalmente, se incorporaron otros objetivos suplementarios tales como obtener datos normativos sobre las medias y otros descriptivos del burnout en la población de deportistas, el cálculo de la consistencia interna y de la homogeneidad de los ítems del ABQ, o el estudio de las posibles diferencias en burnout entre deportistas femeninos y masculinos y entre deportistas de distintos grupos de edad.

MÉTODO

Participantes

La muestra estuvo formada por 320 deportistas pertenecientes a 32 clubes deportivos españoles y 18 deportes diferentes. El 81,6% eran hombres y el 18,4% mujeres, con un rango de edad que oscilaba entre los 12 y los 29 años, siendo la media de 17,03 y la desviación típica de 3,924. La media de entrenamientos semanales se situaba en 3,38, con una desviación típica de 1,537, entrenando un promedio de 1,86 horas por sesión.

Instrumentos de medida

ABQ. Se utilizó la adaptación española realizada por Arce et al. (2010) compuesta por 15 ítems agrupados en 3 factores, con 5 ítems cada uno de ellos: agotamiento físico/emocional (AFE), reducida sensación de logro (RSL) y devaluación de la práctica deportiva (DPD). La escala de respuesta utilizada fue de tipo Likert con 5 alternativas: (1) casi nunca, (2) pocas veces, (3) algunas veces, (4) a menudo y (5) casi siempre.

POMS. Se empleó una versión española reducida del Profile of Mood States (McNair, Lorr y Droppleman, 1971), adaptada al contexto deportivo español por Andrade, Arce, Armental, Rodríguez y de Francisco (2008), con 29 ítems agrupados en 6 factores, dos de ellos positivos (vigor y amistad) y 4 negativos (tensión, depresión, cólera y fatiga). La escala de respuesta utilizada fue tipo Likert con 5 alternativas de respuesta que iban desde Nada (0) a Muchísimo (4).

CSAI-2R. Se utilizó una versión española del Revised Competitive State Anxiety Inventory-2 (Cox, Martens y Russell, 2003), elaborada por Andrade, Lois y Arce (2007). La escala consta de 16 ítems agrupados en 3 factores: ansiedad cognitiva (5 ítems), ansiedad somática (6 ítems) y autoconfianza (5 ítems). La escala de respuesta iba desde (1) Nada a (4) Mucho.

Procedimiento

Se elaboró un cuadernillo que comprendía los tres cuestionarios señalados, en el siguiente orden: POMS, ABQ y CSAI2-R. Los datos fueron recogidos por una psicóloga entre los meses de abril y mayo de 2010. Se siguió un protocolo estandarizado de forma que todos los deportistas recibiesen las mismas instrucciones.

RESULTADOS

El análisis de los datos se realizó mediante el paquete estadístico SPSS. Los resultados se presentan en cuatro apartados que se corresponden con los objetivos de la investigación.

Estadísticos descriptivos de los factores del ABQ

En el factor AFE los deportistas de la muestra obtuvieron una media de 2,0581, en una escala de 1 a 5, siendo la desviación típica ,83226, la asimetría ,858 y la curtosis ,384 . En relación al factor RSL se obtuvo una media de 2,4681, una desviación típica de ,80374, una asimetría de ,306 y una curtosis de -,258. La distribución de las puntuaciones en el componente DPD proporcionó una media de 1,7838, una desviación típica cuyo valor fue ,83571, una asimetría de 1,28 y una curtosis de 1,119. Por último, se obtuvieron las distribuciones de frecuencias a través de las cuales se pudo concluir que tan sólo un 21,9% de los deportistas nunca había experimentado burnout, un 56,9% lo había experimentado pocas veces, un 17,8% algunas veces y un 3,4% lo había experimentado a menudo.

Consistencia interna y homogeneidad de los ítems del ABQ

Para el factor AFE, la consistencia interna medida a través del coeficiente Alpha de Cronbach, fue de ,86, valor que en ningún caso mejoraría si se eliminase un ítem. Para el factor reducida RSL se ha obtenido un valor de Alpha de Cronbach de ,74, cuyo valor no mejoraría si se eliminase cualquiera de sus ítems, El tercer factor DPD mostró una consistencia interna de ,79, la cual se vería incrementada a ,81 si eliminásemos el ítem 15.

Burnout y estados de ánimo

Los coeficientes de correlación de Pearson confirmaron las hipótesis iniciales. Se obtuvieron correlaciones positivas y estadísticamente significativas ($p<0,01$) entre burnout y tensión (,343), depresión (,377), cólera (,281) y fatiga (,361) y correlaciones negativas también estadísticamente significativas ($p<0,01$) entre burnout y vigor (-,351) y entre burnout y amistad (-,125). Se obtuvieron en todos los casos diferencias estadísticamente significativas en la dirección esperada entre los estados de ánimo de los deportistas situados en el cuartil 1 en burnout y los deportistas situados en el cuartil 4. En efecto, para todos los estados de ánimo negativos se han obtenido medias más altas para los sujetos con mayores niveles de burnout y para los estados de ánimo positivos medias más bajas.

Burnout y ansiedad

Los resultados mostraron la existencia de una relación estadísticamente significativa entre burnout y ansiedad somática ($r_{xy} = ,253$, $p < ,01$), pero no se encontró una relación significativa entre burnout y el componente cognitivo de la ansiedad ($r_{xy} = ,107$, $p > ,05$) Se observa que en ambos casos se obtuvieron diferencias estadísticamente significativas entre cuartiles, es decir, mostrando mayores niveles de ansiedad, tanto cognitiva como somática, los sujetos con mayores niveles de burnout.

Burnout y autoconfianza

Se obtuvo una correlación de Pearson entre burnout y autoconfianza de -,429 ($p < ,01$). Se ha obtenido igualmente una diferencia estadísticamente significativa, en la dirección esperada, en la autoconfianza que muestran los sujetos con mayores niveles de burnout y los sujetos con menores niveles. Se observa que los sujetos situados en el cuartil 4 en burnout muestran menores niveles de autoconfianza que los situados en el cuartil 1.

Niveles de burnout por género y grupos de edad

No se encontraron diferencias estadísticamente significativas entre las medias obtenidas por hombres y mujeres en burnout ni en la puntuación total ni en ninguno de sus componentes. Atendiendo a la edad, se hallaron diferencias estadísticamente significativas entre las medias de los distintos grupos de edad y una tendencia ascendente que indica que el burnout tiende a aumentar con la edad de los deportistas.

CONCLUSIONES

Los niveles medios de burnout obtenidos fueron relativamente bajos en los tres factores del ABQ. A pesar de ello, sí se han detectado casos con síndrome de burnout y en riesgo de padecer burnout y se ha obtenido evidencia una vez más de que el burnout no es exclusivo de los deportistas de élite; probablemente esté presente en todos los grupos de edad y en todas las categorías competitivas.

El ABQ ha mostrado índices de consistencia interna aceptables en todos sus factores, situándose en todos los casos el valor de Alpha de Cronbach por encima de ,70.

Las puntuaciones en burnout, obtenidas con el ABQ, han mostrado relaciones directas con los estados de ánimo negativos y con la ansiedad somática pero no con la ansiedad cognitiva. También han mostrado relaciones inversas con los estados de ánimo positivos y con la autoconfianza. Todos estos resultados, con la excepción del obtenido para la ansiedad cognitiva, pueden interpretarse como evidencias de validez externa del ABQ, tanto convergente, dado que lo que mide el ABQ tiene algo en común con estas variables, como discriminante.

No se han encontrado diferencias de género en los niveles de burnout pero sí se han encontrado importantes diferencias entre grupos de edad, aumentando los niveles de burnout a medida que va aumentando la edad de los deportistas; con el aumento de la edad de los deportistas, que suele llevar asociada una mayor carga de práctica deportiva y competición, puede que también aumenten las posibilidades de riesgo de padecer burnout.

NOTA DE LOS AUTORES

La presente investigación ha sido realizada con el apoyo del Ministerio de Ciencia e Innovación y del Fondo Europeo de Desarrollo Regional-FEDER (PSI2010-18807).

REFERENCIAS

Andrade, E., Arce, C., Armental, J., Rodríguez, M. y De Francisco, C. (2008). Indicadores del estado de ánimo en deportistas adolescentes según el modelo multidimensional del POMS. *Psicothema, 20*(4), 630-635.

Andrade, E., Lois, G. y Arce, C. (2007). Propiedades psicométricas de la versión española del Inventario de Ansiedad Competitiva CSAI-2R en deportistas. *Psicothema, 19*(1), 150-155.

Arce, C., De Francisco, C., Andrade, E., Arce, I, y Raedeke, T. (2010). Adaptación española del Athlete Burnout Questionnaire (ABQ) para la medida del burnout en futbolistas. *Psicothema, 22*(2), 250-255.

Cox, R. H., Martens, M. P. y Russell, W. D. (2003). Measuring anxiety in athletics: The Revised Competitive State Anxiety Inventory-2. *Journal of Sport and Exercise Psychology, 25,* 519-533.

De Francisco, C. (2010). *Adaptación psicométrica de una medida de burnout basada en el modelo ABQ de Raedeke y Smith.* Tesis doctoral no publicada, Universidad de Santiago de Compostela, España.

De Francisco, C., Arce, C., Andrade, E., Arce, I. y Raedeke, T. (2009). Propiedades psicométricas preliminares de la versión española del Athlete Burnout Questionnaire en una muestra de jóvenes futbolistas. *Cuadernos de Psicología del Deporte, 9*(2), 45-56.

McNair, D. M., Lorr, M. y Droppleman, L. F. (1971). *Manual for the Profile of Mood States.* San Diego, CA: Educational and Industrial Testing Services.

Raedeke, T. D. y Smith, A. L. (2001). Development and preliminary validation of an athlete burnout measure. *Journal of Sport and Exercise Psychology, 23*(4), 281-306.

CONDUCTA EXPLORATORIA: ANÁLISIS DE DATOS EN ESTUDIANTES UNIVERSITARIOS

Antonio Caballer, Ana Alarcón, Mª José Calero y Aurora Rangel
Universidad Jaume I de Castellón

Resumen

Donoso, Figuera y Torrado (2000) consideran la conducta exploratoria como aquellas acciones cognitivas y conceptuales de la persona que le ayudan a incrementar el conocimiento de sí misma y de su contexto laboral, educativo y social, permitiéndole mejorar la gestión de su proyecto profesional-vital. En el presente trabajo se evalúa la conducta exploratoria en estudiantes universitarios. Para ello, elaboramos una versión de la Escala de Conducta Exploratoria (ESCE) (Donoso, Figuera y Rodríguez, 1994) para estudiantes universitarios (ESCEU) (Caballer y Flores, 2011). Resultados preliminares en universitarios muestran una adecuada estructura factorial para la ESCEU. Las conclusiones más destacables del estudio realizado con 492 estudiantes universitarios de diferentes titulaciones de la Universitat Jaume I de Castellón son que, por una parte, la muestra valora positivamente el prestigio social y la remuneración económica del trabajo, y por otra parte, a pesar de que concede gran importancia a la realización de determinadas actividades relacionadas con la búsqueda de empleo afirma no realizar esas conductas frecuentemente. Además, encontramos diferencias de género estadísticamente significativas en algunos factores relacionados con la conducta de exploración.

A partir de los años sesenta surge la preocupación por definir en qué consiste la exploración del mundo laboral (desde el ámbito personal, educativo y social) y en qué medida esta conducta puede afectar a la elección profesional (Jordaan, 1963). Para Blustein (1992), el proceso exploratorio va más allá de la mera búsqueda de información externa, ya que además de la exploración del entorno (salidas profesionales, conectar con empresas…), también se obtiene información de la exploración de uno mismo, como por ejemplo, pensar acerca de intereses y aptitudes personales. Rodríguez (1999) considera que estos dos procesos están estrechamente relacionados y que no pueden desligarse el uno del otro ya que se influyen mutuamente.

Algunos autores señalan que la conducta de exploración no sólo se relaciona con la elección profesional, sino que además facilita la inserción al trabajo, favorece el ajuste laboral posterior (Blustein, 1988a; Muñoz Sastre, 1992) y se asocia con altos niveles de satisfacción laboral (Donoso, Figuera y Rodríguez, 1994).

En este sentido, estimamos oportuno estudiar la conducta exploratoria una vez que el joven o la joven ya ha decidido formalmente en que rama profesional quiere desarrollar su actividad laboral futura. En nuestro caso concreto, nos planteamos estimar la conducta exploratoria en estudiantes universitarios, ya que, aunque la elección de una carrera universitaria es una de las decisiones más importantes del proceso de la elección profesional, esta decisión no agota todas las acciones que el universitario/a debe llevar a cabo en cuanto a conducta exploratoria se refiere. Por ejemplo, y tal y como señalan Donoso, Figuera y Torrado, (2000), paralelamente a la realización de sus estudios, el universitario puede desarrollar tareas de carrera tales como «identificar vías de inserción» o «generar redes de contacto».

La conducta exploratoria se muestra como un proceso clave en el desarrollo profesional, de ahí la necesidad de desarrollar instrumentos válidos y fiables en la medición de dicho constructo. Con el objetivo de evaluar la conducta exploratoria en jóvenes españoles, Donoso, Figuera y Rodríguez (1994) adaptan la *Career Exploration Survey* (CES) de Blustein (1988b) a muestra española y portuguesa, creando la Escala de Conducta Exploratoria (ESCE). La ESCE ha demostrado ser un instrumento adecuado para evaluar la conducta exploratoria, además de resultar muy útil como apoyo de la acción tutorial (fundamentalmente estudiantes de secundaria y ciclos formativos). La escala utilizada en la presente investigación es una versión de la ESCE para universitarios ya que el objetivo principal de este trabajo es ampliar la evaluación de la conducta exploratoria a estudiantes universitarios de la Universitat Jaume I que se encuentran en proceso de transición al mercado laboral. Otro objetivo que nos proponemos es comprobar la posible existencia de diferencias de género en las distintas facetas que conforman la conducta exploratoria.

MÉTODO

Participantes

492 estudiantes universitarios de nueve titulaciones diferentes de la Universitat Jaume I de Castelló. 205 eran hombres y 287 eran mujeres cuya media de edad era de 20,87 años (D.T.=12,30).

Instrumento

Escala de Conducta Exploratoria en Universitarios (ESCEU) (Caballer y Flores, 2011). Formada por 42 ítems tipo Likert de 5 alternativas de respuesta, es una adaptación para universitarios de la Escala de Conducta Exploratoria (ESCE). La consistencia interna de la ESCEU, calculada con el alfa de Cronbach, es de 0,86. Esta escala ha mostrado una adecuada estructura factorial en muestra universitaria, así como coeficientes de consistencia interna adecuados para todos los factores

(entre 0.70 y 0.82). Concretamente se obtienen 9 factores: METAS (importancia que se le concede a determinadas actividades dirigidas a conseguir metas laborales), EXPECTATIVAS (expectativas relativas a encontrar trabajo), SEGURIDAD (confianza que tiene el estudiante sobre la información laboral que dispone), FUTURO (actividades llevadas a cabo por el estudiante para preparar su futuro profesional), SATISFACCIÓN (satisfacción sobre la información laboral obtenida), ANSIEDAD (ansiedad ante determinadas situaciones profesionales), INFORMACIÓN (información relacionada con el proceso general de conseguir un trabajo), REFLEXIÓN (proceso de reflexión personal sobre el rendimiento académico y futuro profesional) y PRESTIGIO (importancia que se le concede al prestigio profesional y a su remuneración económica).

Procedimiento

La aplicación de la escala se llevó a cabo en horas lectivas de las distintas titulaciones seleccionadas.

Resultados y conclusiones

Como se observa en la tabla 1, la muestra valora positivamente el prestigio social y la remuneración económica (Prestigio), además de conceder gran importancia a la realización de determinadas actividades relacionadas con la búsqueda de empleo (Metas). También concede importancia al proceso de reflexión personal sobre el futuro profesional. Sin embargo, no realiza conductas (Futuro) dirigidas a la consecución de dichas dimensiones. Nuestros resultados son congruentes con Rodríguez (1999) quien manifiesta que existen dos fuentes de donde se obtiene información acerca del proceso de exploración: el entorno y uno mismo. En nuestro caso concreto, nuestra muestra de estudiantes concede mayor importancia a la reflexión (uno mismo) que a la acción. A la luz de estos resultados, resultaría adecuado desarrollar programas de intervención destinados a que los jóvenes inicien y/o aumenten actividades dirigidas a la obtención de información del entorno, y no sólo la proveniente de uno mismo. Como es ampliamente aceptado, las creencias y percepciones, reales o no, tienen un efecto sobre la conducta (Ajzen, 1991; Fishbein y Ajzen, 1975). En el caso concreto de la conducta exploratoria, una de las creencias que adquiere especial relevancia es la percepción que se tiene sobre el mercado de trabajo. De este modo, las personas estarán menos dispuestas a invertir esfuerzos si consideran que su acción no resultará provechosa, por lo que el nivel de implicación en la exploración será menor (Donoso, Figuera y Rodríguez, 1994). En este sentido, en futuros trabajos cabe plantearse si la actual crisis económica y laboral está teniendo algún efecto y en qué medida sobre la exploración profesional por parte de los jóvenes, máxime cuando algunos autores señalan que la conducta exploratoria, además de facilitar la exploración también favorece salir de situaciones de desempleo continuado y el ajuste laboral posterior (Donoso, Figuera y Rodríguez, 1994).

Tabla 1. Medias de las dimensiones de la ESCEU

Dimensiones	Medias	D.T
Prestigio	3.77	0.85
Metas	3.63	0.73
Reflexión	3.46	0.89
Expectativas	3.34	0.82
Ansiedad	3.22	0.80
Seguridad	2.92	0.87
Información	2.85	0.79
Satisfacción	2.56	0.77
Futuro	1.86	0.73

Además, tal y como muestra la tabla 2, encontramos diferencias de género estadísticamente significativas en algunas dimensiones relacionadas con la conducta de exploración, concretamente en Reflexión, Metas y Ansiedad. Según nuestros resultados, las mujeres realizan más actividades dirigidas a alcanzar objetivos vocacionales que los hombres, a la vez que dicen reflexionar más que los hombres sobre su futuro académico y profesional. También son las mujeres las que obtienen medias más altas en ansiedad. Es necesario llevar a cabo más investigaciones para intentar dilucidar el sentido de esas diferencias, ya que, por ejemplo, no podemos concluir si la mayor ansiedad manifestada por las mujeres es una reacción a una mayor exploración o viceversa.

Como ya comentamos, entendemos la conducta exploratoria como un proceso continuado en el tiempo, y no como una decisión puntual que se reduce a la etapa de la adolescencia. Por ese motivo, en este trabajo nos hemos planteado la necesidad de ampliar la evaluación de la conducta exploratoria a estudiantes universitarios. Consideramos que también podría resultar interesante ir más allá y abordar la tarea de evaluar dicha conducta una vez que la persona forma parte de la red laboral ya que, probablemente, la exploración de ese entorno puede facilitar la elaboración de proyectos profesionales más realistas.

Tabla 2. Diferencias de medias por género

Dimensiones	Media Hombres	Media Mujeres	t
Reflexión	3.34	3.53	-2,27*
Metas	3.53	3.71	-2,77**
Ansiedad	3.12	3.29	-2,32*

** p < .05, ** p < .01*

REFERENCIAS

Ajzen, I. (1991). The theory of planned behaviour. *Organizational Behavior and Human Decision Processes, 50 (2*), 179-211.

Blustein, D.L. (1988a). Individual and contextual factors in career exploration. *Journal of Vocational Behavior, 33*, 203-216.

Blustein, D.L. (1988b). The relationship between motivational processes and career exploration. *Journal of Vocational Behavior, 33*, 340-357.

Blustein, D. L. (1992). Applying current theory and research in career exploration to practice. *Career Development Quartely, 41*, 174-184.

Caballer, A. y Flores, R. (2011). Análisis de una escala para medir la conducta exploratoria en estudiantes universitarios. *International Journal of Developmental and Educational Psychology, 1*(2), 61-70.

Donoso, T., Figuera, P. y Rodríguez, M.L. (1994). Un instrumento para evaluar la conducta exploratoria en el desarrollo de la carrera profesional. *Revista de Investigación Educativa, 23*, 490-495.

Donoso, T., Figuera, P. y Torrado, M. (2000). Análisis y validación de una escala para medir la conducta exploratoria. *Revista de Investigación Educativa, 18*(1), 202-220.

Fishbein, M. y Ajzen, I. (1975). *Belief, attitude, intention and behavior: an introduction to theory and research*. Reading. Massachusetts: Addison-Wesley.

González, V. (2009). Autodeterminación y conducta exploratoria. Elementos esenciales en la competencia para la elección profesional responsable. *Revista Iberoamericana de Educación, 51*, 201-220.

Jordaan, J. P. (1963). Exploratory behavior: The formation of self and occupational concepts. En D. Super, R. Starishevsky, N. Matlin y J. P. Jordaan (Eds.), *Career development: Self-concept theory: Essays in vocational development* (pp. 42-78). New York: College Entrance Examination Board.

Muñoz Sastre, M. T. (1992). Calidad de la información profesional y «disponibilidad» de índices sociales. *Revista de Psicología General y Aplicada, 45(*2), 161-167.

Rodríguez, M. L. (1999). Enseñar a explorar el mundo del trabajo. *Diagnóstico de las destrezas exploratorias y propuestas de intervención*. Málaga: Ediciones Algibe.

Rodríguez, M.L., Sandín, Mª P. y Buisán, C. (2000). La conducta exploratoria: concepto y aplicaciones en Orientación Profesional. *Revista de Educación*, 321, 153-186.

Super, D. E. (1994). A life-span, life-space perspective on convergence. En M. L. Savickas y R. L. Lent (Eds.), *Convergence in career development theories* (pp. 63-74). San Francisco: Jossey-Bass.

Taveira, M. (1997). Exploraçao e desenvolvimento vocacional de jovens, Estudio sobre as relaçoes entre a exploraçao, a identidade e a indecisao vacacional. Tesis doctoral no publicada, Universidade do Minho, Braga, Portugal.

VALIDEZ CONVERGENTE DE DOS INSTRUMENTOS PARA LA MEDIDA DEL BURNOUT EN DEPORTISTAS

Cristina de Francisco[1], Enrique Garcés de los Fayos[2], Mar Graña[1], Iria Arce[1] y Constantino Arce[1]

[1] Universidad de Santiago de Compostela

[2] Universidad de Murcia

Correo electrónico: constantino.arce@usc.es

Resumen

La finalidad principal de este estudio fue la de analizar la validez convergente y discriminante de la adaptación española del Athlete Burnout Questionnaire (ABQ). Con este objetivo, una muestra de 460 deportistas de ambos sexos con edades entre los 13 y los 34 años, cubrieron un cuadernillo que contenía el ABQ y otra herramienta de medida del burnout en deportistas, el Inventario de Burnout en Deportistas (IBD). El ABQ evalúa agotamiento físico y emocional (AFE), reducida sensación de logro (RSL) y devaluación de la práctica deportiva (DPD). Por su parte, el IBD mide agotamiento físico (AE), reducida realización personal (RRP) y despersonalización (D). Ambos instrumentos miden el burnout en deportistas a través de tres dimensiones, existiendo entre ellas cierta equivalencia. Los resultados obtenidos a través de los análisis multi-rasgo/multi-método muestran cierto grado de convergencia entre dos medidas del burnout en deportistas, entre las dimensiones AFE y AE, por un lado, y entre RSL y RRP por otro. Respecto a DPD y D, no se han observado indicios de convergencia.

Este estudio se ha realizado con el objetivo de aportar pruebas sobre la validez convergente y discriminante de la adaptación española del Athlete Burnout Questionnaire (ABQ) mediante la aproximación multi-rasgo/multi-método de Campbell y Fiske (1959). Para ello se aplicó este instrumento junto con el Inventario de Burnout en Deportistas (IBD) a una muestra de deportistas españoles de diferentes disciplinas. Tanto el ABQ como el IBD son herramientas ideadas para la medición del síndrome de burnout en deportistas, pero tienen sus diferencias. El ABQ se basa en una conceptualización tridimensional del síndrome caracterizada por el agotamiento físico y emocional (AFE) de los deportistas, la reducida sensación de logro (RSL) y la devaluación de la práctica deportiva (DPD). El IBD evalúa igualmente el burnout en deportistas como un constructo tridimensional, cuyos componentes son el agotamiento emocional (AE), la reducida realización personal (RRP) y la despersonalización (D). Hipotéticamente, deberían encontrarse correlaciones más altas entre los rasgos teóricamente equivalentes medidos por ambos cuestionarios (AFE y AE, RSL y RRP) que entre distintos rasgos (AFE, RSL y DPD en el ABQ

y AE, RRP y D en el IBD) medidos por el mismo cuestionario. Además, si realmente son distintos los rasgos DPD del ABQ y D del IBD, debería encontrarse una correlación baja entre ambos y, en todo caso, más baja que entre AFE y AE y entre RSL y RRP. Además de los dos cuestionarios señalados, se ha administrado a los deportistas un cuestionario más para la medida de la satisfacción deportiva, con dos rasgos: satisfacción/diversión (S/D) y aburrimiento (A). Bajo hipótesis, la satisfacción y el aburrimiento deberían correlacionar negativa y positivamente, de forma respectiva, con todos los rasgos del burnout, tanto del ABQ como del IBD.

MÉTODO

Participantes

La muestra estuvo formada por 460 deportistas españoles (247 hombres y 213 mujeres) de diferentes modalidades y niveles deportivos. Sus edades oscilaban entre los 13 y los 34 años (media = 18.94; desviación típica = 2.93). El número de sesiones de entrenamiento semanales se situaba entre 1 y 11 (media= 3.51; desviación típica = 1.56), y las horas diarias de entrenamiento iban desde 1 a 7 (media = 2.05; desviación típica = 0.91).

Instrumentos

Versión española del Athlete Burnout Questionnaire (Arce, De Francisco, Andrade, Arce y Raedeke, 2010; De Francisco, 2010). El cuestionario consta de 15 ítems, divididos equitativamente en tres sub-escalas para la medida de las dimensiones del burnout propuestas por Raedeke (1997): agotamiento físico y emocional, reducida sensación de logro y devaluación de la práctica deportiva.

Inventario de Burnout en Deportistas (Garcés de Los Fayos, 1999). Se utilizó una versión reducida de 19 ítems para evaluar las tres dimensiones del burnout propuestas por Maslach y Jackson (1981): agotamiento emocional, reducida realización personal y despersonalización.

Versión española del Cuestionario de Satisfacción Intrínseca en el Deporte, abreviadamente SSI (Balaguer, Atienza, Castillo, Morena y Duda, 1997). El cuestionario se compone de 8 ítems distribuidos en dos sub-escalas que evalúan la satisfacción/diversión en la práctica deportiva (S/D) con 5 ítems y el aburrimiento (A) con 3 ítems. En este estudio se aplicó una versión de 7 ítems (se prescindió de un ítem en la dimensión aburrimiento) en base a los resultados obtenidos en la investigación de Castillo, Balaguer y Duda (2002).

Procedimiento

Se diseñó un cuadernillo que contenía los instrumentos de medida señalados y una sección adicional para datos personales. Se utilizó un procedimiento de reco-

gida de datos estandarizado de forma que todos los deportistas recibieran las mismas instrucciones.

RESULTADOS

En la Tabla 1 se ofrece la matriz multi-rasgo/multi-método con las correlaciones entre los rasgos respectivos de los tres instrumentos de medida utilizados en la investigación y los coeficientes Alpha de Cronbach en la diagonal principal. Las correlaciones se han calculado mediante un Análisis Factorial Confirmatorio realizado con el programa LISREL, donde se especificaron 8 factores, 3 del ABQ, 3 del IBD y 2 del SSI. De acuerdo con una de las hipótesis se observa como las correlaciones más altas se producen entre AFE y AE (r_{xy}= .71, p<.01) y entre RSL y RRP (r_{xy}= .49, p<.01), mismos rasgos medidos por distintos métodos; y también, de manera esperada, se observa como la correlación más baja se produce entre DPD y D (r_{xy}=.02, p>.05) que, al no alcanzar significatividad estadística, indica que, en efecto, se trata de rasgos diferentes. También se puede observar en la Tabla que los resultados ofrecen, al mismo tiempo, cierto grado de evidencia de la validez discriminante del ABQ siendo las correlaciones homo-rasgo superiores a las correlaciones hetero-rasgo con la excepción de la correlación entre RSL y DPD (r_{xy}= .57, p<.01) que supera a la correlación entre RSL y RRP (r_{xy}= .49, p<.01).

Respecto a la relación entre los distintos rasgos del burnout y la satisfacción deportiva (S/D), se observan en la mayoría de los casos, de acuerdo con las hipótesis, correlaciones negativas que oscilan entre -.38 y -.58 para el ABQ y entre -.02 y -.62 para el IBD, todas ellas estadísticamente significativas (p<.01), con la excepción de la correlación entre D y S/D. Mientras que con el aburrimiento (A) ocurre justamente lo contrario; es decir, se han obtenido correlaciones positivas, estadísticamente significativas en ambos cuestionarios (p<.01), con la excepción de la correlación entre D y A.

Tabla 1. Matriz multi-rasgo/multi-método obtenida mediante AFC con los 8 factores

		ABQ			IBD			SSI	
		AFE	**RSL**	**DPD**	**AE**	**RRP**	**D**	**S/D**	**A**
ABQ	AFE	.83							
	RSL	.36	.68						
	DPD	.29	.57	.81					
IBD	AE	.71	.62	.58	.73				
	RRP	.23	.49	.43	.42	.74			
	D	.18	.13*	.02*	.27	-.01*	.74		
SSI	S/D	-.38	-.51	-.58	-.62	-.52	-.02*	.79	
	A	.42	.43	.61	.65	.53	.07*	-.91	.59

Nota: Todas las correlaciones alcanzaron significatividad estadística (p<.01) excepto las marcadas con *. Los valores de la diagonal principal son coeficientes Alpha de Cronbach. Por último, se calcularon las correlaciones entre las puntuaciones totales de los tres cuestionarios utilizados en la investigación obteniéndose una correlación positiva de .59 (p<.01) entre las puntuaciones totales del ABQ y el IBD, de -.50 (p<.01) entre IBD y SSI, y de -.56 (p<.01) entre ABQ y SSI.

CONCLUSIONES

Los resultados de la presente investigación permiten concluir que tan sólo dos de los rasgos del ABQ (AFE y RSL) convergen, respectivamente con dos de los rasgos del IBD (AE y RRP). Respecto al tercer rasgo medido por ambos cuestionarios (DPD en ABQ y D en IBD) no se ha alcanzado convergencia, lo que indica que se trata de rasgos conceptualmente diferentes. Sí convergen, no obstante, los totales de ambos cuestionarios y los rasgos del ABQ con los rasgos de la satisfacción deportiva: diversión y aburrimiento.

NOTA DE LOS AUTORES

La presente investigación ha sido realizada con el apoyo del Ministerio de Ciencia e Innovación y del Fondo Europeo de Desarrollo Regional-FEDER (PSI2010-18807).

REFERENCIAS

Arce, C., De Francisco, C., Andrade, E., Arce, I. y Raedeke, T. (2010). Adaptación española del Athlete Burnout Questionnaire (ABQ) para la medida del burnout en futbolistas. *Psicothema, 22*(2), 250-253.

Balaguer, I., Atienza, F. L., Castillo, I., Moreno, Y. y Duda, J. L. (1997). Factorial structure of measures of satisfaction/interest in sport and classroom in the case of Spanish adolescents.*Abstracts of 4*th. *European Conference of Psychological Assessment* (p. 76). Lisbon: Portugal.

Campbell, D. T. y Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Castillo, I., Balaguer, I y Duda, J. L. (2002). Las perspectivas de meta de los adolescentes en el contexto deportivo. *Psicothema, 14*(2), 280-287.

De Francisco, C. (2010). *Adaptación psicométrica de una medida de burnout basada en el modelo ABQ de Raedeke y Smith*. (Tesis doctoral; Recurso electrónico). Santiago de Compostela: Servizo de Publicacións e Intercambio Científico, Universidade de Santiago de Compostela.

Garcés de Los Fayos, E. J. (1999). *Burnout en deportistas: un estudio de la influencia de variables de personalidad, sociodemográficas y deportivas en el síndrome*. Tesis Doctoral no publicada, Universidad de Murcia, Murcia.

Maslach, C. y Jackson, S. E. (1981). *MBI: Maslach Burnout Inventory. Manual.* Palo Alto, CA: Consulting Psychologists Press.

Raedeke, T. D. (1997). Is athlete burnout more than just stress? A sport commitment perspective. *Journal of Sport and Exercise Psychology, 19*(4), 396-417.

ACTITUDES ANTE LA MUERTE EN PACIENTES DE VIH/SIDA

Antonio, López-Castedo e Isabel Calle Santos
Universidad de Vigo
Correo electrónico: alopez@uvigo.es

Resumen

Los sentimientos ante las circunstancias claves de la vida y, en particular, ante la muerte se van construyendo como resultado de un proceso de sus experiencias y creencias y, en especial, en función de las actitudes generadas a lo largo de la vida. El propósito del presente trabajo fue analizar la consistencia interna y la validez de constructo de la Escala Multidimensional de Miedo a la Muerte (EMMM) en sujetos de VIH/SIDA. La escala se ha aplicado a una muestra de 148 sujetos de ambos sexos y con edades comprendidas entre los 23 y 50 años. El coeficiente alfa de consistencia interna ha sido de .71. El análisis factorial exploratorio presenta una estructura de 8 factores que explican el 46,22% de la varianza. Los resultados ponen de manifiesto que la escala resulta particularmente útil en la población de interés, posee adecuadas propiedades psicométricas y posibilitando desarrollar una intervención psicológica en la mejora de la calidad de vida en sujetos con VIH/SIDA.

A mitad de los años 60, cuando Templer comenzó a investigar sobre el concepto de ansiedad ante la muerte, ésta era un tema tabú para los científicos del comportamiento y los profesionales de la salud mental (Templer, 1970). Desde entonces muchas cosas han cambiado, y ahora es un campo de investigación fértil, sin embargo no ha habido una teoría que lo abarcara todo, aunque el enfoque basado en la teoría de los constructos personales de Kelly (1955) y centrado en la estructura cognitiva de la ansiedad ante la muerte (Epting y Neimeyer, 1984; Neimeyer, 1988) es bastante prometedor.

Las actitudes (Llor, Abad, García & Nieto, 1995) se definen como el conjunto de creencias que los individuos tienen sobre los objetos específicos de la realidad, en este caso, la muerte. Implican una combinación de conceptos, informaciones y emociones, que predisponen a los sujetos a responder de un modo favorable o desfavorable a personas, grupos o ideas del medio social u objetos concretos del entorno físico. Las actitudes son parte de las estructuras cognitivas que la persona desarrolla para organizar y sistematizar sus experiencias y conductas. Las actitudes proporcionan una guía simplificada y práctica que indica cuáles son las conductas apropiadas y ayudan a la persona a prever y dirigir los eventos periódicos.

Los elementos fundamentales son: a.- *Componente cognitivo*, es el factor informativo o conceptual, formado por el conjunto de creencias que la persona tiene sobre el objeto de la actitud. b.- *Componente evaluativo o emocional*: implica una valoración del objeto en la dimensión positiva como atractivo o la negativa como indeseable. Implica una respuesta emocional ante el objeto actitudinal. c.- *Componente conductual*: es el elemento motivacional de la actitud que englobaría la intención conductual de responder ante el objeto, en función de la aceptación ó rechazo.

Actualmente, la perspectiva existencial de que las personas necesitan tener una sensación de significado para sobrevivir y afrontar la muerte ofrece un marco conceptual útil para integrar varios patrones de actitudes hacia la muerte.

El presente estudio tiene como objetivo el de exponer las propiedades psicométricas (consistencia interna y validez de constructo) de la Escala Multidimensional de Miedo a la Muerte (EMMM) en el ámbito clínico.

MÉTODO

Participantes

La muestra total está compuesta por 148 sujetos infectados de VIH y pacientes enfermos de SIDA de la provincia de Ourense, con una media de de edad de 35.57 (DT = 5.76

Instrumento

Se administró la Escala Multidimensional de Miedo a la Muerte (EMMM) de Hoelter (1972) y Walkey (1982). Consta de 42 ítems con un formato de respuesta tipo Likert de 5 puntos (1 = Totalmente de acuerdo, 5 = Totalmente en desacuerdo). Evalúa 8 dimensiones: miedo al proceso de morir, a los muertos, a ser destrozado, por otras personas significativas, a lo desconocido, a una muerte consciente, por el cuerpo después de la muerte, y a una muerte prematura. La fiabilidad como consistencia interna oscila entre .65 a .85. La validez con otras escalas abarca desde .49 a .70.

Procedimiento

Una vez obtenida la autorización y conformidad de las diferentes instancias, los sujetos completaron la prueba de forma colectiva, voluntaria y anónima. Dada la peculiaridad de la muestra se aprovecharon los distintos seminarios mensuales para pasarle la prueba, siendo administrada por el mismo investigador.

RESULTADOS

Análisis de los ítems

En la tabla 1 se presentan los estadísticos descriptivos del cuestionario, junto con el índice de homogeneidad corregido, observando que todos los ítems que conforman la Escala Multidimensional de Miedo a la Muerte mantienen una correlación elevada con la puntuación total de la escala, con un rango que oscila entre .23 (ítem 17) y .79 (ítem 3). Todos los ítems fueron significativamente homogéneos con un margen de error del uno por mil.

Tabla 1. Media (M), Desviación Típica (DT), Indice de Homogeneidad corregido (IH) y Alpha excluido el ítem del EMMM

Item	Media	D.T.	IH-1	α-1	Item	Media	D.T.	IH-1	α-1
1	1.54	1.28	.35	.78	**22**	1.33	1.02	.38	.78
2	4.09	1.54	.45	.77	**23**	3.22	1.84	.52	.77
3	4.31	1.46	**.79**	.77	**24**	1.20	.84	.34	.78
4	1.51	1.29	.35	.77	**25**	2.56	1.76	.33	.78
5	3.76	1.79	.46	.76	**26**	3.34	1.88	.35	.77
6	2.93	1.81	.41	.78	**27**	2.02	1.65	.41	.77
7	4.46	1.34	46	.77	**28**	4.42	1.39	.28	.79
8	2.44	1.81	.45	.77	**29**	1.78	1.55	.30	.77
9	1.53	1.21	.41	.78	**30**	4.01	1.64	.47	.77
10	1.71	1.48	.45	.78	**31**	1.70	1.49	.44	.77
11	4.55	1.18	.43	.77	**32**	1.61	1.37	.30	.79
12	2.98	1.91	.47	.77	**33**	4.26	1.49	.72	.77
13	1.40	1.13	.44	.77	**34**	1.15	.61	.31	.78
14	3.78	1.77	.36	.78	**35**	4.45	1.36	.57	.77
15	3.95	1.70	.62	.77	**36**	1.09	.57	.30	.78
16	1.40	.99	.41	.77	**37**	2.49	1.48	.35	.78
17	1.45	.96	**.23**	.78	**38**	2.24	1.74	.40	.77
18	3.62	1.83	.55	.76	**39**	4.48	1.26	.41	.77
19	2.47	1.89	.40	.78	**40**	4.40	1.36	.43	.77
20	4.53	1.21	.41	.77	**41**	2.39	1.82	.27	.78
21	2.40	1.81	.57	.77	**42**	2.29	1.83	.41	.77

Fiabilidad

La consistencia interna se ha calculado mediante el índice Alfa de Cronbach (Tabla 2). En términos generales, se pone de manifiesto un índice alto y con una tendencia uniforme en las ocho dimensiones que evalúa la Escala Multidimensional de Miedo a la Muerte.

Tabla 2. Fiabilidad: consistencia interna (alfa de Cronbach) del EMMM

FACTORES		F1	F2	F3	F4	F5	F6	F7	F8
GENERO	*Hombres*	.62	.57	.77	.56	.63	.58	.63	.64
	Mujeres	.66	.58	.60	.57	.62	.59	.71	.72
LUGAR DE RESIDENCIA	*Urbana*	.66	.59	.75	.57	.63	.61	.62	.69
	Semiurbana	.68	.60	.67	.59	.63	.64	.65	.68
	Rural	.64	.68	.80	.60	.69	.66	.67	.67
PERCEPCIÓN SALUD	*Buena*	.64	.70	.69	.59	.69	.65	.69	.70
	Regular	.66	.65	.77	.58	.64	.61	.69	.73
	Mala	.68	.63	.75	.59	.65	.59	.65	.69
ENFERMEDAD ACTUAL	*Sida*	.71	.68	.70	.61	.65	.62	.70	.71
	Vih	.70	.70	.76	.69	.68	.69	.70	.73
TRATAMIENTO	*Si*	.69	.68	.73	.68	.69	.61	.69	.72
	No	.65	.66	.75	.69	.68	.67	.68	.70
VIA DE CONTAGIO	Parenteral	.70	.66	.74	.69	.68	.69	.65	.60
	Sexual	.71	.69	.80	.68	.70	.70	.70	.73
	No sabe	.79	.67	.71	.70	.69	.67	.71	.74
TOTALES		**.70**	**.69**	**.77**	**.68**	**.70**	**.69**	**.71**	**.72**

Análisis factorial exploratorio

Se ha realizado un análisis de componentes principales con rotación Oblimin. El índice de Kaiser-Meyer-Olkin (KMO) que ofrece un valor de .801, estimado como bueno e indicando que las correlaciones entre parejas de ítems pueden ser explicadas por los restantes ítems seleccionados, y la prueba de esfericidad de Bartlett ($\chi^2 = 1718{,}08$; gl = 861; $p<.001$,) mostrando que los ítems no eran independientes, por lo que se garantiza que el análisis factorial es adecuado y el modelo consigue un buen ajuste. El Determinante obtenido de las correlaciones fue de .004. Del análisis se extrajeron ocho factores que explican el 46,22 de la varianza (tabla 3).

Los factores pueden ser identificados como:

Factor 1: Miedo a una muerte prematura, compuesto por cinco ítems que reflejan una preocupación por que la muerte pueda impedir que se cumplan metas importantes en la vida o tener experiencias significativas y por cierto miedo a la inexistencia, a lo desconocido ($\alpha = 73$).

Factor 2: Miedo a ser destrozado, compuesto por tres ítems que incluyen la disecación y cremación del cuerpo ($\alpha = 85$).

Factor 3: Miedo al proceso de morir, formado por nueve ítems en la que se exhiben muertes dolorosas y violentas ($\alpha = 69$).

Tabla 3. Análisis Factorial de Componentes principales con rotación OBLIMIN del EMMM

	Factores								
Item	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**h^2**
21	.71								.56
12	.69								.51
8	.66								.49
18	.60								.55
5	.40								.46
3		.86							.78
33		.85							.75
15		.77							.65
10			.69						.53
13			.67						.58
31			.67						.49
22			.51						.48
1			.48						.37
27			.46						.47
42			.40						.40
29			.36						.34
24			.33						.33
35				.66					.67
40				.64					.53
7				.58					.51
20				.55					.45
9				.53					.42
30				.41					.48
32				.40					.28
28				.32					.27
2					.68				.56
23					.66				.50
14					.60				.45
26					.48				.35
25					.47				.31
39					.36				.37
37						.57			.35
16						.54			.49
17						.51			.32
34						.44			.23
4						.43			.35
41						.41			.34
6							.71		.55
19							.63		.46
38							.33		.40
36								.63	.43
11								.55	.57
Varianza	5.09	2.85	2.62	2.17	1.91	1.69	1.60	1.47	
% Var.	**12.11**	**6.79**	**6.25**	**5.17**	**4.55**	**4.02**	**3.82**	**3.51**	
% Acum.	12.11	18.90	25.14	30.32	34.87	38.90	42.71	46.22	

Factor 4: Miedo por el cuerpo después de la muerte, integrado por ocho ítems que expresan preocupación por la decadencia y el aislamiento del cuerpo ($\alpha = 61$).

Factor 5: Miedo a los muertos, constituido por seis ítems que reflejan evitación de restos humanos y animales ($\alpha = 64$).

Factor 6: Miedo por otras personas significativas, formado por seis ítems que revelan aprensión respecto al impacto de la muerte del sujeto en otras personas significativas y de los muertos de dichas personas en el sujeto ($\alpha = 61$).

Factor 7: Miedo a una muerte consciente, integrado por tres ítems que manifiestan ansiedad respecto a ser declarado muerto falsamente ($\alpha = 65$).

Factor 8: Miedo a lo desconocido, constituido por dos ítems que hacen referencia a un miedo por el cuerpo después de la muerte ($\alpha = 60$).

DISCUSIÓN

A tenor de los resultados obtenidos podemos afirmar que las propiedades psicométricas del EMMM son satisfactorias y corroboran las investigaciones previas (Holter, 1979; Walkey, 1982), ofreciendo una medida consistente y refinada de un amplio espectro de miedos a la muerte. Los índices de Homogeneidad son adecuados, revelando una buena cohesión entre los respectivos ítems y los valores de consistencia interna pueden considerarse aceptables.

En general se extrae de las investigaciones sobre las Actitudes ante la Muerte, y particularmente desde una perspectiva existencial, que las personas tienen la necesidad de buscar significado a su vida y es la ausencia de significado lo que producirá más miedo a la muerte, que está en línea con lo propuesto por autores como Lewis y Butler (1974) que sugieren que cuando las personas ven sus vidas como plenas y significativas, deberían mostrar menos ansiedad y más aceptación ante la muerte. Por su parte, Erikson (1963), propone que los sujetos que conciben su vida con un sentido de integridad, considerando que ha valido la pena y que han acercado la distancia existente entre la realidad y el ideal, tienen mayores posibilidades de afrontar la muerte sin miedo. Durlack (1972), subraya que los sujetos con menos miedo a la muerte, con actitudes más positivas y aceptación de la misma, eran los que consideraban a la vida con propósito y le daban significado; a su vez, Flint, Gayton y Ozmon (1983), encontraron correlación positiva en una dirección equivalente a los anteriores, es decir cuando existía satisfacción con el propio pasado, la aceptación de la muerte era mayor.

En síntesis, las actitudes de las personas hacia la muerte tienden a ser una mezcla, en las que hay actitudes conflictivas que contrabalancean las otras para adaptarse (Feifel, 1990). La incertidumbre y alta tasa de letalidad asociada al VIH/SIDA constituye un motivo más que suficiente para justificar esta investigación, con el fin de localizar los factores implicados en estos pacientes y llevar a cabo intervenciones modificadoras. Por tanto, en la educación sobre la muerte y el asesoramien-

to psicológico, es importante tener presente, por un lado, la multidimensionalidad de la persona (social, espiritual, somática, psicológica) funcionando como un sistema conjunto, integrado, dinámico y fluctuante, y, por otro, el carácter dinámico y flexible (variable en el tiempo) en función de la evolución de la enfermedad, las prioridades, los valores, las necesidades, entre otras.

REFERENCIAS

Durlack, J. (1972. Relationship between individual attitudes toward life and death. *Journal of Consulting and Clinical Psychology*, 38, 463-469.

Epting, F. R., y Neimeyer, R. A. (1984). *Personal meanings of death: Applications of personal construct theory to clinical practice*. New York: Hemisphere/McGraw Hill.

Erikson, E. (1963). *Childhood and society* (2ª ed.). New York: Norton.

Feifel, H. (1990). Psychology and death. *American Psychologist*, 45, 537-543.

Flint, G.A., Grayton, W.F. y Ozmon, K.L. (1983). Relationships between life satisfaction and acceptance of death by elderly persons. *Psychological Reports.* 53, 290-292.

Hoelter, J. W. (1979). Multidimensional treatment of fear of death. *Journal of Consulting and Clinical Psychology*, 47, 996-999.

Lewis, M.I. y Butler, R.N. (1974). Life-review therapy: Putting memories to work in individual and group psychoterapy. *Geriatrics*, 29 (11), 165-173.

Llor, B., Abad, M.A., García, M. y Nieto, J. (1995). *Ciencias Psicosociales Aplicadas a la Salud. Aspectos psicosociales de la hospitalización* (227-246). Madrid: McGraw-Hill.

Neimeyer, R. A. (1988). Death anxiety. En H. Wass, F. M. Berardo, y R.A. Neimeyer (Comps.): *Dying: Facing the facts* (97-136). Washington, DC: Hemisphere.

Templer, D.I. (1970). The construction and validation of a death anxiety scale. *Journal of General Psychology,* 82, 165-177.

Walkey, F.H. (1982). The Multidimensional Fear of Death Scale: An independent analysis. *Journal of Consulting and Clinical Psychology*, 50, 466-467.

EVALUACION PSICOMETRICA DEL GHQ-30 EN ADOLESCENTES

Antonio, López-Castedo y José Domínguez Alonso
Universidad de Vigo
Correo electrónico: alopez@uvigo.es

Resumen

La adolescencia es un período crítico del ciclo de la vida que está lleno de cambios bio-psico-sociales, en el cual la evaluación del bienestar psicológico es un aspecto importante tanto para la investigación como para la intervención psicosocioeducativa. El presente estudio tiene como objetivo analizar la adecuación y funcionamiento del Cuestionario de Salud General de 30 ítems (GHQ-30) en la población adolescente. La muestra, compuesta por 2.083 estudiantes de ambos sexos de Educación Secundaria Obligatoria (2º ciclo) y Bachillerato, atesora un rango de edad entre los 15 y 19 años. El análisis de fiabilidad mostró una alta consistencia interna ($\alpha = .93$) y estabilidad temporal ($r = .89$). El análisis factorial exploratorio identificó cuatro factores (Ansiedad/Insomnio, Depresión severa, Adaptación inadecuada, Disfunción social) que explican el 49,96% de la varianza. Se concluye que el Cuestionario GHQ-30 posee adecuadas propiedades psicométricas y resulta expresamente útil en la población de interés con un peso relevante en la evaluación de determinados aspectos del malestar psicológico.

La adolescencia constituye un importante período de cambios biopsicológicos, mentales, afectivos y sociales, siendo además, una etapa de transición en el curso del desarrollo humano, al implicar el paso progresivo de la infancia a la edad adulta (Arnett, 2000; Dávila, 2004; Musitu, Buelga, Lila y Cava, 2001).

La investigación científica de las últimas décadas evidencia el enorme impacto que esta etapa tiene en la salud mental (Bolognini, Plancherel, Betts-Chart y Halfon, 1996; Eccles, 1999; Sweeting y West, 2003). Estudios recientes revelan la importancia de factores psicológicos (ansiedad, estrés, depresión) en el comportamiento de los adolescentes, poniendo en grave riesgo su salud (Aalto-Setaelae, Marttunen, Tuulio-Henriksson, Poikolainen y Loennqvist, 2002; Benjet y Hernandez Guzman, 2001; Mahon, Yarcheski y Yarcheski, 2003; Weiss et al., 2005). Por ello, resulta relevante disponer en la práctica cotidiana de instrumentos con calidad métrica contrastada para la detección de síntomas ansiosos, depresivos, somáticos y de dificultades psicosociales que están presentes en esta fase del ciclo vital.

El Cuestionario de Salud Mental (G.H.Q.) de Goldberg (1978), en sus diferentes versiones de 60, 30, 28 y 12 ítems, ha sido ampliamente utilizado para evaluar la salud autopercibida, especialmente en lo que se refiere a la presencia de síntomas emocionales (Werneke, Goldberg, Yalcin y Ustun, 2000). Según Goldberg & Williams (1988), el cuestionario fué concebido para analizar el funcionamiento de dos áreas: en primer lugar, la capacidad de realizar actividades sanas más usuales y, en segundo lugar, la aparición de síntomas estresantes nuevos. Aunque la mayor parte de las investigaciones se han basado en poblaciones de adultos (Goldberg et al., 1997), también se ha utilizado con éxito en adolescentes (French y Tait, 2004; Sweeting, Young y West, 2009).

En esta línea la presente investigación tiene como objetivo analizar la homogeneidad de los ítems, la fiabilidad como criterio métrico de la calidad global (consistencia interna y estabilidad temporal) y la estructura factorial exploratoria del G.H.Q. en su versión de 30 ítems. En concreto se aplica el cuestionario a una muestra de adolescentes que se encuentran en una edad especialmente propicia de alteraciones psíquicas provocadas por cualquier tipo de modificaciones físicas, psíquicas y/o sociales.

MÉTODO

Participantes

En esta investigación han participado 2.083 estudiantes de ambos sexos de Educación Secundaria Obligatoria (2º ciclo) y Bachillerato, siendo la edad media de 15,90 (DT = 2,10). Los sujetos pertenecían a centros públicos y concertados de Galicia.

Intrumento

El G.H.Q. utilizado ha sido el de 30 ítems (adaptación y validación española de Lobo y Muñoz, 1991). La respuesta a cada pregunta se puntuó según el método de Likert en una escala de cuatro puntos (0-1-2-3). La fiabilidad como consistencia interna oscila entre .82 a .93 y como estabilidad temporal entre .80 a .82. La validez con otras escalas abarca desde .45 a .77.

Procedimiento

Una vez obtenida la autorización y conformidad de las diferentes instancias, los estudiantes completaron la prueba de forma colectiva, voluntaria y anónima. La estrategia básica de aplicación consistió en administrar el G.H.Q.-30 en dos ocasiones con un período de tiempo de 20 días. Los investigadores fueron siempre los mismos.

RESULTADOS

Análisis de los ítems

La tabla 1 presenta los estadísticos descriptivos del cuestionario, junto al índice de homogeneidad corregido, observando que todos los ítems que conforman el GHQ-30 mantienen una correlación elevada con la puntuación total de la escala, y un rango que oscila entre .27 (ítem 5) y .70 (ítem 23). Todos los ítems fueron significativamente homogéneos con un margen de error del uno por mil.

Tabla 1. Media (M), Desviación Típica (DT), Indice de Homogeneidad corregido (IH) y Alpha excluido el ítem del GHQ -30

Item	Media	D.T.	IH-1	α-1	Item	Media	D.T.	IH-1	α-1
1	1.09	.58	.49	.93	**16**	1.02	.81	.60	.92
2	.82	.78	.50	.93	**17**	1.03	.63	.53	.93
3	.71	.75	.53	.93	**18**	.92	.75	.58	.92
4	.91	.53	.32	.93	**19**	.55	.67	.52	.93
5	1.02	.71	**.27**	.93	**20**	.95	.56	.55	.93
6	.99	.42	.40	.93	**21**	.76	.81	.69	.92
7	.98	.53	.52	.93	**22**	.79	.81	.68	.92
8	.97	.60	.54	.93	**23**	.51	.73	**.70**	.93
9	.81	.57	.29	.93	**24**	.43	.69	.63	.93
10	.85	.58	.42	.93	**25**	.43	.67	.61	.92
11	.86	.62	.39	.93	**26**	.86	.64	.54	.93
12	.96	.60	.54	.93	**27**	.91	.62	.55	.93
13	.89	.55	.46	.93	**28**	.69	.79	.66	.92
14	.86	.82	.66	.92	**29**	.43	.73	.62	.93
15	.79	.78	.69	.92	**30**	.58	.74	.63	.93

Fiabilidad

Para evaluar la fiabilidad del GHQ-30 se ha calculado la consistencia interna y la estabilidad temporal (tabla 2). La consistencia interna se determina mediante el Coeficiente Alpha de Cronbach (Cronbach, 1951). Se alcanzan índices de fiabilidad muy elevados con una tendencia uniforme tanto en el género como en el curso escolar, superando el criterio de .70 recomendado por Nunnaly & Berstein (1994). En la estabilidad temporal, se observan correlaciones altas y significativas, obteniéndose en la muestra total una correlación de .89.

Tabla 2. Fiabilidad: consistencia interna (alfa de Cronbach) y estabilidad temporal (test-retest) del GHQ-30

VARIABLES	METODO	
	α Cronbrach	**Test - retest**
SEXO		
Mujeres	.94	.90
Hombres	.91	.89
CURSO		
3º Eso	.92	.89
4º Eso	.94	.90
1º Bachiller	.93	.90
2º Bachiller	.93	.89
TOTAL	**.93**	**.89**

ANÁLISIS FACTORIAL EXPLORATORIO

La validez factorial del cuestionario GHQ-30 se evaluó mediante el análisis factorial exploratorio (AFE) de los ítems, utilizando la técnica de extracción de ejes principales y el método de rotación varimax. El índice de Kaiser-Meyer-Olkin (KMO) ofrece un valor de .959 -estimado como muy bueno-, prediciendo que las correlaciones entre parejas de ítems pueden ser explicadas por los restantes ítems seleccionados. La prueba de esfericidad de Bartlett (χ^2 = 25008,68; gl = 435; p<.001) muestra que los ítems no eran independientes, por lo que garantiza un análisis factorial adecuado y un modelo que consigue un buen ajuste. El Determinante obtenido en las correlaciones fue de .006.

Del análisis se extrajeron cuatro factores que explican el 49,98% de la varianza (tabla 3), con saturaciones que oscilan entre .39 y .76, criterio a partir del cual son consideradas como buenas (Comrey, 1973).

El primer factor explica el 34,45% de la varianza común y compuesto por 12 ítems que hacen referencia a un estado de desazón, nerviosismo, miedo a lo desconocido, inestabilidad y dificultades para conciliar el sueño. Se denomina «*Ansiedad/Insomnio*» (α = .90). El segundo factor constituido por seis ítems, explica el 7,09% de la varianza y refleja los estados anímicos de la persona (tristeza profunda, inseguridad, sentimiento de inferioridad, ideación suicida) que afectan a la actividad diaria: visión negativa del mundo, pérdida del interés de hacer algo y de disfrutar la vida. Se etiquetó como «*Depresión severa*» (α = .85). El tercer factor explica el 4,70% de la varianza y está formado por ocho ítems que expresan la incapacidad de tomar decisiones, de concentrarse y de ver el futuro con esperanza. Se denomina «*Adaptación inadecuada/Inhibición*» (α = .79). Finalmente, el cuarto factor explica el 3,73% de la varianza. Integrado por 4 ítems plantea cómo la persona se siente en la relación con otras personas ante los problemas y dificultades que se enfrenta. Se denomina «*Disfunción social*» (α = .61).

Tabla 3. Análisis Factorial de Componentes principales con rotación VARIMAX del GHQ-30

	Factores				
Item	**F1**	**F2**	**F3**	**F4**	**h2**
2	.71				.57
14	.70				.60
21	.65				.59
3	.64				.46
18	.62				.47
15	.61				.57
16	.60				.48
22	.60				.56
19	.58				.42
28	.56				.55
30	.50				.51
17	.39				.41
24		.76			.69
25		.74			.65
29		.69			.61
23		.63			.63
26		.54			.51
12		.47			.47
8			.68		.55
7			.65		.50
20			.55		.45
6			.54		.34
13			.53		.41
1			.51		.38
4			.49		.31
27			.43		.40
11				.75	.62
10				.64	.52
9				.60	.42
5				.51	.32
Varianza	10.33	2.13	1.41	1.12	
% Var.	34.45	7.09	4.70	3.73	
% Acum.	34.45	41.54	46.24	49.98	

Las correlaciones entre el GHQ- 30 y los factores (tabla 4) sustentan la elección de la rotación factorial realizada, ya que sus valores son todos estadísticamente significativos ($p < .001$) y elevados.

Tabla 4. Correlación entre el GHQ-30 y los factores

	GHQ- 30 T	F1 (Ansiedad/ Insomnio)	F 2 (Depresión Severa)	F3 (Adaptación inadecuada/ Inhibición)	F4 (Disfunción Social)
F1	***.93***				
F2	***.85***	***.71***			
F3	***.82***	***.63***	***.62***		
F4	***.57***	***.38***	***.37***	***.52***	

DISCUSIÓN

A tenor de los resultados obtenidos podemos afirmar que el GHQ en su versión de 30 ítems ha demostrado excelentes propiedades psicométricas en una muestra de adolescentes. El análisis de homogeneidad y las correlaciones de cada ítem con la puntuación de la escala corregida (de moderadas a elevadas con valores de r hasta .70), indican una correcta construcción del cuestionario con respecto a sus contenidos y confirman su utilidad, relevancia y conveniencia para su aplicación.

En relación a su fiabilidad, la consistencia interna (Alpha de Cronbach) es muy satisfactoria, con tendencia bastante uniforme en relación al sexo y curso escolar, atestiguando que, efectivamente, el rendimiento en nuestro medio es muy similar al documentado en otros estudios (Chan, 1985; Chan y Chan, 1983; Goodchild y Duncan-Jones, 1985; Keyes, 1984; Shek, 1987). En el análisis de la estabilidad temporal (test-retest), se obtienen altas correlaciones que muestran una elevada reproducibilidad cuando se administra el cuestionario bajo las mismas condiciones, en dos ocasiones separadas por 20 días (De Paulo y Folstein, 1978).

La validez factorial exploratoria, evidencia un modelo compuesto por cuatro factores, confirmando que los factores extraídos representan adecuadamente las dimensiones teóricas. Los datos obtenidos, revelan cierta similitud con las investigaciones llevadas a cabo por Shek, 1987 e Iwata, Uno y Susuki, 1994, aunque con matices debido a la utilización de muestras distintas. A su vez, los coeficientes alpha de Cronbach, siempre mayores de .60, presentan suficiente consistencia interna y sugieren que los ítems de cada factor miden un constructo unitario con escaso error aleatorio.

En conclusión, las características de brevedad y facilidad de administración del cuestionario y la constatación de adecuadas propiedades psicométricas y validez de constructo, permiten disponer de un instrumento apropiado para la evaluación e investigación sobre la salud mental o el bienestar psicológico en adolescentes, tanto en contextos de investigación como clínicos. En definitiva, el cuestionario del GHQ-30 se muestra como una medida precisa, fiable y válida a través de culturas y del tiempo, proponiendo una validación confirmatoria de los resultados obtenidos y la determinación de su sensibilidad, aspectos éstos en los que se sigue trabajando.

REFERENCIAS

Aalto-Setaelae, T., Marttunen, M., Tuulio-Henriksson, A., Poikolainen, K., y Loennqvist, J. (2002). Depressive symptoms in adolescence as predictors of early adulthood depressive disorders and maladjustment. *American Journal of Psychiatry, 159,* 1235–1237.

Arnett, J.J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychlogist*, 55 (5), 469-480.

Benjet, C. y Hernandez, L. (2001). Gender differences in psychological well-being of mexican early adolescents. *Adolescence,* 36 (141), 47-65.

Bolognini, M., Plancherel, B., Betts-Chart, W. y Halfon, O. (1996). Self-esteem and mental health in early adolescence: Development and gender differences. *Journal of Adolescence*, 19, 233-245.

Chan D.W. y Chan T.S. (1983). Reliability, validity and the structure of the General Health Questionnaire in a Chinese context. *Psychological Medicine*, 13, 363-371.

Chan, D.W. (1985). The Chinese version of the General Health Questionnaire: does language make a difference? *Psychological Medicine,* 15, 147-155.

Comrey, A.L. (1973). *A first course in factor analysis*. New York: Academic Press.

Cronbach, J.L. (1951). Coefficient alpha and internal structure of a test. *Psychometrika*, 16, 297-334.

Dávila, O. (2004). Adolescencia y juventud: De las nociones a los abordajes. *Última Década*, 21, 83-104.

De Paulo, J.R. y Folstein, M.F. (1978). Psychiatric disturbances in neurological patients: detection, recognition, and hospital course. *Annals of Neurology,* 4, 225-228.

Ecles, J. S. (1999). The Development of children ages 6 to 14.*The Future of Children*, 9 (2), 30-44.

French, D.J. y Tait, R.J. (2004). Measurement invariance in the General Health Questionnaire-12 in young Australian adolescents. *European Child and Adolescent Psychiatry*, 13, 1-7.

Goldberg, D.P. (1978). *Manual of the General Health Questionnaire*. Oxford: NFER-NELSON.

Goldberg, D. P. y Williams, P. (1988). *A user´s guide to the General Health Questionnaire*. Windsor, UK: The NFER-NELSON.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O. y Rutter, C. (1997). The validity of two version of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine*, 27, 191-197.

Goodchild, M.E. y Duncan-Jones P. (1985). Chronicity and the General Health Questionnaire. *British Journal of Psychiatry*, 146, 55-61.

Iwata, N., Uno, B. y Susuki, T. (1994). Psychometric properties of the 30-ítem General Health Questionnaire in Japanese. *Psychiatry and Clinical Neurociences*, 48 (3), 547-556.

Keyes, S. (1984). Gender stereotypes and personal adjusment: employing the PAQ, TSBI and GHQ with samples of British adolescents. *British Journal of Social Psychology*, 23, 173-180.

Lobo, A. y Muñoz, P.E. (1996). *Cuestionario de Salud General GHQ (General Healh Questionnaire). Guía para el usuario de las distintas versiones en lengua española validadas*. Barcelona: Masson.

Mahon, N. E., Yarcheski, A. y Yarcheski, T. J. (2003). Anger, anxiety and depression in early adolescents from intact and divorced families. *Journal of Pediatric Nursing*, 18 (4), 267-273.

Musitu, G., Buelga, S., Lila, M. y Cava, M. J. (2001). *Familia y adolescencia*. Madrid: Síntesis.

Nunnaly, J.C. y Berstein, I.H. (1994). *Psychometric theory (3ª ed.)*. New York: McGraw-Hill.

Shek, D. T. (1987). Reliability and factorial structure of the Chinese version of the Genral Health Questionnaire. *Journal of Clinical Psychology,* 43, 683-691.

Sweeting, H. y West, P. (2003). Sex differences in health at ages 11, 13 and 15. *Social Science and Medicine*, 56, 31-39.

Sweeting, H., Young, R., y West, P. (2009). GHQ increases among Scottish 15 year olds 1987-2006. *Social Psychiatry and Psychiatric Epidemiology*, 44, 579-586.

Weiss, J. W., Mouttapa, M., Chou, C. P., Nezami, E., Andersson, C., Palmer, P.H., Cen, S., Gallaher, P., Ritt-Olson, A., Azen, S. y Unger, J.B. (2005). Hostility, depressive symptoms, and smoking in early adolescence. *Journal of Adolescence*, 28, 49-62.

Werneke, U., Goldberg, D. P., Yalcin, Y. y Ustun, B. T. (2000). The stability of the factor structure of the General Health Questionnaire. *Psychological Medicine*, 30, 823-829.

PROPIEDADES PSICOMÉTRICAS DE LA VERSIÓN ESPAÑOLA DEL *SOCIAL PROVISIONS SCALE (SPS)* EN UNA MUESTRA DE ESTUDIANTES UNIVERSITARIOS

Zeltia Martínez, María Soledad Rodríguez, María Adelina Guisande, Carolina Tinajero y María Fernanda Páramo
Universidad de Santiago de Compostela

Resumen

El *Social Provisions Scale* (SPS; Cutrona y Russell, 1987) es una medida de apoyo social percibido que evalúa las diferentes funciones de las relaciones sociales. El presente estudio tiene como objetivo analizar las propiedades psicométricas de las puntuaciones de la versión española del SPS. La escala se aplicó a una muestra de 855 estudiantes (575 mujeres y 280 hombres) de primer año de grado de la Universidad de Santiago de Compostela. Los resultados del análisis factorial confirman la estructura de seis factores propuesta por los autores en la versión original. La consistencia interna de la puntuación total y de las seis subescalas fue satisfactoria, así como las evidencias de validez convergente obtenidas de la relación con la escala de apoyo social SSQ6. Finalmente, el análisis de las puntuaciones en función del género mostró que las mujeres perciben más apoyo que los hombres. Los resultados indican que la versión española del SPS es una medida adecuada para evaluar el apoyo social percibido en población universitaria, por lo que puede ser aplicada con suficientes garantías en estudios sobre logro académico y ajuste a la universidad.

El apoyo social percibido se ha convertido en uno de los factores de protección más importantes durante la transición a la universidad (Pratt et al., 2000). El apoyo social percibido es un constructo multidimensional conceptualizado como la valoración cognitiva que tiene el estudiante de que puede comunicarse con los otros, que es valorado, aceptado, protegido y de que se le ayudará cuando lo necesite (Sarason, Sarason, y Pierce, 1990).

La teoría de las provisiones sociales elaborada por el sociólogo Robert Weiss (1974) ha contribuido de forma significativa a la conceptualización del apoyo social. Weiss reconoce que los diferentes contextos, problemas o situaciones vividas por los individuos requieren diferentes formas de apoyo que denominó provisiones sociales y que identificó como: *alianza confiable, orientación, apego, integración social, refuerzo de valía y oportunidad de cuidar*. La alianza confiable y la orientación hacen referencia a la seguridad o certeza de que se puede contar con los otros

para la asistencia tangible y el consejo/ información, respectivamente. La provisión de apego representa la proximidad emocional y sensación de seguridad con los otros. La integración social se centra en el sentimiento de pertenencia a un grupo con el que se comparten preocupaciones, intereses y actividades. El refuerzo de valía es el reconocimiento de los otros de nuestra propia competencia, aptitudes y valores, y por último, la oportunidad de cuidar refleja la sensación que uno tiene de que es necesario para los otros. Las provisiones de alianza confiable y orientación se caracterizan por contribuir directamente a la resolución de problemas en momentos de estrés (función instrumental), mientras que las otras destacan por promover el bienestar del individuo independientemente de los niveles de estrés (función emocional). Para Weiss, cada una de estas seis provisiones se obtiene normalmente de un tipo particular de relación social, aunque múltiples provisiones pueden ser obtenidas de una misma persona. En cualquier caso, para que una persona se sienta apoyada es necesario que todas las provisiones estén presentes en sus interacciones personales.

A partir de la operativización de las seis provisiones sociales de Weiss, Carolyn Cutrona y Daniel Russell (1987) elaboran el *Social Provisions Scale* (SPS) un instrumento que evalúa la percepción individual de las funciones del apoyo social, a través del grado de presencia o ausencia de cada provisión en las relaciones interpersonales que mantiene la persona. Las propiedades psicométricas del SPS, obtenidas en muestras de estudiantes universitarios norteamericanos (Cutrona, Cole, Colangelo, Assouline, y Russell, 1994; Cutrona y Russell, 1987), canadienses (Caron, 1996), portugueses (Moreira y Canaipa, 2007) y pakistaníes (Rizwan y Syed, 2010), indican que esta escala constituye una medida excelente para evaluar el apoyo social percibido en el contexto de la Educación Superior. El SPS ha mostrado una estructura de seis factores relacionados, coherente con las seis provisiones sociales de Weiss, y coeficientes de consistencia interna adecuados para la puntuación total de la escala y las puntuaciones de las subescalas. Asimismo, numerosos estudios, la mayoría realizados con estudiantes universitarios, han encontrado evidencias de su validez convergente y discriminante (Caron, 1996; Cutrona et al., 1994; Moreira y Canaipa, 2007; Rizwan y Syed, 2010; Sherry, Law, Hewitt, Flett, y Besser, 2008; Wintre y Yaffe, 2000).

El presente estudio tiene como objetivo analizar las propiedades psicométricas del SPS en una muestra de estudiantes universitarios españoles, en relación a su estructura factorial, consistencia interna y correlación con otras variables.

MÉTODO

Participantes

La muestra estaba compuesta por 855 universitarios (575 mujeres, 280 hombres) de primer año de grado de 16 titulaciones de la Universidad de Santiago de Compostela, con edades comprendidas entre los 17 y los 20 años.

Instrumentos

Versión española del *Social Provisions Scale* (SPS; Cutrona y Russell, 1987). Consta de 24 ítems con un formato de respuesta Likert de 4 puntos (1=totalmente en desacuerdo, 4= totalmente de acuerdo), que evalúan las seis provisiones del apoyo social percibido enunciadas por Weiss (1974): alianza confiable, orientación, apego, integración social, refuerzo de valía y oportunidad de cuidar. Cada provisión es medida por cuatro ítems, dos evalúan la presencia de la provisión y los otros dos su ausencia.

Versión española del *Social Support Questionnaire- Short Form* (SSQ6; Sarason, Sarason, Shearin, y Pearcy, 1987). Consta de 6 ítems divididos cada uno de ellos en dos partes. La primera evalúa el número de personas que el sujeto percibe como disponibles en caso de necesidad (SSQ6N). La segunda evalúa el grado de satisfacción con el apoyo percibido (SSQ6S) en una escala Likert de seis puntos (1=muy insatisfecho, 6= muy satisfecho).

Procedimiento

En un estudio previo, los ítems de la escala fueron traducidos y administrados a una muestra piloto de 277 universitarios. Atendiendo a criterios cuantitativos y cualitativos, algunos de estos ítems fueron revisados y reformulados. La escala derivada de este proceso es la que se empleó en este estudio.

Para la recogida de datos se elaboró un cuaderno que incluía un cuestionario sobre características sociodemográficas y las versiones adaptadas al español del SPS y el SSQ6. Los estudiantes completaron los cuestionarios, de forma voluntaria y anónima, en el aula.

RESULTADOS

Análisis Factorial Confirmatorio

Para analizar la estructura de seis factores propuesta por Cutrona y Russell (1987) en la versión original, se llevó a cabo un análisis factorial confirmatorio utilizando el programa LISREL 8.8 (Jöreskog y Sörbom, 2006). Debido a la falta de normalidad multivariada de los datos, el método de estimación empleado fue el de máxima verosimilitud robusto (Satorra y Bentler, 1994). Las cargas factoriales fueron significativas, poniendo de manifiesto que los ítems son relevantes para definir sus correspondientes dimensiones. El ajuste del modelo fue evaluado a través de diferentes índices: Chi-cuadrado de Satorra-Bentler (χ^2_{SB}), la razón entre χ^2_{SB} y sus grados de libertad ($\chi^2_{SB}/g.l.$), el índice de ajuste comparativo (CFI), el índice de bondad de ajuste (GFI), el error cuadrático medio de aproximación (RMSEA) y la raíz media cuadrática residual (RMR). Los valores observados en los

índices pusieron de manifiesto que la estructura de seis factores presenta un buen ajuste a los datos: χ^2_{SB} = 660.64 (g.l.= 237), $\chi^2_{SB}/g.l.$ = 2.78, CFI= .98, GFI= .91, RMSEA= .046 y RMR= .025.

Fiabilidad

El valor del coeficiente de fiabilidad alfa de Cronbach fue de .874 para la puntuación total de la escala. Para las subescalas alcanzó valores de .751 (alianza confiable), .731 (orientación), .678 (integración social), .644 (apego), .597 (refuerzo de valía) y .545 (oportunidad de cuidar).

Relación entre las puntuaciones del SPS y del SSQ6

En la tabla 1 se recogen los coeficientes de correlación de Pearson entre las dimensiones del SPS, así como las correlaciones entre el SPS y la medida de apoyo social SSQ6.

Tabla 1. Correlaciones entre las dimensiones del SPS y del SSQ6

	SPSTotal	AC	Or	Ap	IS	RV	OC
Alianza Confiable	.760*						
Orientación	.806*	.772*					
Apego	.779*	.478*	.603*				
Integración social	.727*	.528*	.504*	.439*			
Refuerzo de valía	.699*	.395*	.429*	.458*	.400*		
Oportunidad de cuidar	.658*	.360*	.354*	.369*	.387*	.371*	
SSQ6 Disponibilidad	.392*	.321*	.317*	.272*	.304*	.290*	.246*
SSQ6 Satisfacción	.326*	.300*	.335*	.278*	.200*	.151*	.198*

* $p < .01$.

Las seis provisiones del apoyo social presentaron relaciones positivas y significativas entre sí, lo que indica que cuando un estudiante dispone de algún tipo de provisión social también cuenta con el resto de funciones de las interacciones personales. La correlación más elevada fue entre las provisiones de orientación y alianza confiable ($r=.772$, $p<.05$), ambas dirigidas a la satisfacción de necesidades de tipo instrumental. Las correlaciones más bajas se obtuvieron para oportunidad de cuidar, que representa la única función del apoyo en la cual el individuo es el proveedor y no el receptor de la ayuda.

Las relaciones obtenidas entre las puntuaciones del SPS y el SSQ6 fueron todas directas y significativas, con un rango de valores entre .151 y .392, por lo que un estudiante que tiene acceso a todas las provisiones sociales, dispone de una mayor red de apoyo social y está más satisfecho con el apoyo percibido.

Diferencias en las dimensiones del SPS en función del género

Atendiendo a las diferencias en las dimensiones del SPS en función del género, observamos que las puntuaciones de las mujeres fueron significativamente más altas que las de los hombres tanto en la percepción de apoyo global ($t= 4.26$, $p<.00$), como en las provisiones de apego ($t= 4.49$, $p<.01$), integración social ($t= 2.24$, $p<.05$), oportunidad de cuidar ($t= 0.16$, $p<.01$), alianza confiable ($t= 2.26$, $p<.05$) y orientación ($t= 3.56$, $p<.01$).

DISCUSIÓN

El objetivo del presente estudio fue analizar las propiedades psicométricas de las puntuaciones del SPS en una muestra española de universitarios de primer año de grado, atendiendo a su estructura factorial, consistencia interna y relación con otras variables.

Los resultados derivados del análisis factorial confirmatorio coinciden con la estructura de seis factores relacionados propuesta por Cutrona y Russell (1987) a partir del modelo teórico de las provisiones sociales de Robert Weiss: alianza confiable, orientación, apego, integración social, refuerzo de valía y oportunidad de cuidar.

La consistencia interna de las dimensiones, analizada mediante el coeficiente alfa de Cronbach, fue similar a la obtenida en la escala inglesa así como en las adaptaciones al francés y al urdu (Caron, 1996; Cutrona y Russell, 1987; Rizwan y Syed, 2010).

Respecto a la relación entre las seis provisiones sociales, los resultados obtenidos ponen de manifiesto que cuando una persona tiene acceso a una forma de apoyo, también tiene acceso a los demás funciones (Caron, 1996; Cutrona y Russell, 1987; Moreira y Canaipa, 2007). Aquellos estudiantes que perciben que son queridos y valorados disponen de una red social con la que compartir intereses y actividades, recursos materiales e información. La correlación más elevada se obtuvo entre las dos provisiones de tipo instrumental, alianza confiable y orientación, y las más débiles fueron para oportunidad de cuidar.

Por otro lado, las provisiones del SPS presentan una relación positiva con las facetas del apoyo social percibido que evalúa el SSQ6, disponibilidad y satisfacción con la red de apoyo. Al igual que en investigaciones previas (Cutrona y Russell, 1987; Moreira y Canaipa, 2007), se constata que cuanto mayor es el grado de presencia de las provisiones sociales en las relaciones del joven, mayor es la red social de apoyo disponible y la satisfacción con la ayuda percibida.

En cuanto a las diferencias en función del género, se observa que las mujeres experimentan mayores niveles de apoyo social que los hombres. De acuerdo con Vaux (1985), estas diferencias pueden ser explicadas por los patrones de socialización tradicionales. Entre las mujeres se fomentan actividades vinculadas al cuida-

do, afiliación y expresión de emociones, mientras que entre los hombres se prescriben la autonomía e independencia.

La versión española del SPS es una medida adecuada para la evaluación del apoyo social percibido en población universitaria. En el contexto de la Educación Superior, disponer de un instrumento de medida como el SPS permitirá obtener el perfil de las fortalezas y debilidades de las funciones del apoyo social en las interacciones de los jóvenes universitarios, facilitando el diseño e implementación de acciones de asesoramiento y orientación psicológica dirigidas a los estudiantes que presenten problemas de adaptación en su primer año de ingreso en la universidad.

REFERENCIAS

Caron, J. (1996). L'échelle de provisions sociales: Une validation québécoise. *Santé Mentale au Québec, 21*, 158-180.

Cutrona, C., Cole, V., Colangelo, N., Assouline, S., y Russell, D.W. (1994). Perceived parental social support and academic achievement: An attachment theory perspective. *Journal of Personality and Social Psychology, 66*, 369-378.

Cutrona, C., y Russell, D. (1987). The provisions of social support and adaptation to stress. En W.H. Jones y D. Perlman (Eds.), *Advances in personal relationships, Vol.1*, (pp.37-67). Greenwich, CT: JAI Press.

Jöreskog, K., y Sörbom, D. (2006). *LISREL 8.80.* Chicago: Scientific Software.

Moreira, J.M., y Canaipa, R. (2007). A Escala de Provisões Sociais: Desenvolvimento e validação da versão portuguesa da «Social Provisions Scale». *Revista Iberoamericana de Diagnóstico y Evaluación Psicológica, 24*, 23-58.

Pratt, M., Hunsberger, B., Pancer, S., Alisat, S., Browers, C., Mackey, K., Ostaniewicz, A., Rog, E., Tezzian, B., y Thomas, N. (2000). Facilitating the transition to university: Evaluation of Social Support Discussion Intervention Program. *Journal of College Student Development, 41,* 427-441.

Rizwan, M., y Syed, N. (2010). Urdu translation and psychometric properties of Social Provision Scale. *The International Journal of Educational and Psychological Assessment, 4,* 33-47.

Sarason, B.R., Sarason, G.R., y Pierce, G.R. (1990). *Social support: An interactional view*. New York: John Wiley and Sons.

Sarason, I.G., Sarason, B.R., Shearin, E.N., y Pierce, G.R. (1987). A brief measure of social support: Practical and theoretical implications. *Journal of Social and Personal Relationships, 4,* 497- 510. doi: 10.1177/0265407587044007

Satorra, A., y Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. En A.von Eye y C.C. Clogg (Eds.), *Latent variable analysis in developmental research* (pp. 285–305). Thousand Oaks, CA: Sage.

Sherry, S., Law, A., Hewitt, P., Flett, G., y Besser, A. (2008). Social support as a mediator of the relationship between perfectionism and depression: A preliminary test of the social disconnection model. *Personality and Individual Differences, 45*, 339-344. doi:10.1016/j.paid.2008.05.001

Vaux, A. (1985). Variations in social support associated with gender, ethnicity, and age. *Journal of Social Issues, 41,* 89-110. doi: 10.1111/j.1540-4560.1985.tb01118.x

Weiss, R. (1974). The provisions of social relations. En Z. Rubin (Ed.), *Doing unto others* (pp. 17-26). Englewood Cliffs: Prentice- Hall.

Wintre, M.G., y Yaffe, M. (2000). First-year students´ adjustment to university life as a function of relationships with parents. *Journal of Adolescent Research, 15,* 9-37. doi: 10.1177/0743558400151002

ASSESSMENT OF BODY IMAGE: PSYCHOMETRIC PROPERTIES OF THE QÜIC IN SPANISH ADOLESCENT GIRLS AND BOYS

Eva Penelo, Paola Espinoza, Mariona Portell y Rosa M. Raich
Universidad Autónoma de Barcelona
Correo electrónico: eva.penelo@uab.cat

Abstract

This study investigated the psychometric properties of the Body Image Questionnaire (QÜIC) in Spanish adolescents. The sample comprised 254 girls and 189 boys, aged 12-15. Principal component analyses showed that the 18 satisfaction items could be summarized using two moderately interrelated dimensions, torso and head/limbs, with satisfaction with chest/breast and genitals loading on a different factor for boys (torso) and girls (head/limbs). The QÜIC measures of body satisfaction, body problems, general physical appearance, and conformity with weight and height presented satisfactory test-retest reliability, internal consistency, and convergent and discriminant validity. Our findings support the use of the QÜIC when assessing body image.

The altered perception of, and dissatisfaction with, body image are among the most common features of patients with eating disorders (ED), but also occur in the general population (e.g., Neumark-Sztainer, Paxton, Hannan, Haines, & Story, 2006). Its assessment in adolescents is especially important because during adolescence there are significant physical and psychological changes in self-image that play a unique role in the construction of identity and gender role. Moreover, since the DSM-IV-TR (APA, 2000) includes over-concern with weight and shape as a criterion for the diagnosis of anorexia nervosa and bulimia nervosa, it is essential to properly assess this ED component. Several self-report questionnaires have been developed (Thompson & van den Berg, 2002), most of them using whole schematic silhouettes of the body to be rated globally. However, it may be better to examine specific body parts (Stanford & McCabe, 2002).

The body image construct includes an affective/evaluative component (self-ideal discrepancies and body satisfaction-dissatisfaction) and a cognitive/investment component (importance of cognitive-behavioral salience of one's appearance) (Cash, Melnyk, & Hrabosky, 2004). The literature often neglects body image investment, but it is important to study both dimensions (Giovannelli, Cash, Henson, & Engle, 2008), since each of them can have a different effect on eating disturbances (Allen, Byrne, McLean, & Davis, 2008). Specifically, cognitive/invest-

ment dimension appears to be more powerful in predicting eating attitudes and behaviors (Cash et al., 2004; Espinoza, Penelo, & Raich, 2010).

In this scenario, we present the validation of the Spanish version of the QÜIC (*Qüestionari d'Imatge Corporal*) in a sample of adolescent girls and boys.

METHOD

Participants and Procedure

A total of 254 girls and 189 boys (age: *M*=13.5 years; *SD*=0.4) from seven schools in Terrassa (Catalonia, Spain) were included by means of incidental sampling. Weight status was: 4.8% underweight, 63.4% normal-weight, 27.3% overweight, and 4.5% obesity.

The study was approved and mediated by the Terrassa Municipal Institute of Health and Social Welfare. Consent from parents and adolescents were obtained. The questionnaires (see below) were voluntarily answered in class, as part of a larger assessment. Statistical analyses were conducted with SPSS v.19 for Windows.

Instruments

Qüestionari d'Imatge Corporal (QÜIC; Miró, 2006). The QÜIC was initially developed in Catalan (Miró, 2006) from clinical experience to assess body image aspects considered in the DSM in children and adolescents. It contains three sections: a) satisfaction (affective dimension) with each of the 18 parts of the body that appear in a drawing of a girl or a boy (0-10; scores derived from the average of the corresponding items), and whether each of these parts of the body constitutes a problem (cognitive dimension) (*yes*/*no*; scores obtained through the sum of items endorsed); b) general physical appearance (0-10), the higher, the better; and c) conformity with current weight and height (*conformity*, *wish for more*, or *wish for less*).

Eating Attitudes Test (EAT-40; Garner & Garfinkel, 1979). This 40-item self-report questionnaire assesses attitudes, feelings and behaviors characteristic of ED. We applied the Spanish adaptation (Castro, Toro, Salamero, & Guimerà, 1991). The total score we used showed satisfactory internal consistency (Cronbach's α=.86).

Cuestionario de Influencias del Modelo Estético Corporal (CIMEC-26: Toro, Castro, Gila, & Pombo, 2005; CIMEC-V: Toro, Salamero, & Martínez, 1994). This 26-item self-report questionnaire evaluates the impact that different social agents can have on the development of attitudes to one's body in males and females aged 12-24. In the present sample the internal consistency was satisfactory (Cronbach's α=.92).

Eating Disorder Examination Questionnaire (EDE-Q; Fairburn & Beglin, 1994). It includes 22 attitudinal items on ED psychopathology over the past 28 days. We used the Shape Concern (EDE-Q-SC) score (8 items) of the Spanish ad-

aptation (Penelo, Villarroel, Portell, & Raich, in press; Villarroel, Penelo, Portell, & Raich, 2011), which showed satisfactory internal consistency in our sample (Cronbach's α=.94).

RESULTS

Factor Structure and Internal Consistency

Body satisfaction items were analyzed separately for each sex, using principal component analysis, with oblimin-oblique rotation. Listwise deletion was applied. Only components with an eigenvalue >1 were retained and the Cattell's scree test was applied. Items showing cross-loading were allocated to the factor with the highest loading, when the difference with respect to the second highest value (in absolute value) was above .10. Otherwise, the contribution of the item to the internal consistency of each scale was examined, based on *Cronbach's α if item deleted* coefficient.

Table 1. Body satisfaction: Principal components analysis

	Girls (*n*=254)			**Boys (*n*=189)**		
% explained variance	52.1			53.5		
(KMO)	(.88)			(.90)		
Factor loadings	**F1**	**F2**	h^2	**F1**	**F2**	h^2
Satisfaction with						
hair	.00	**.60**	*.37*	−.15	**.78**	*.48*
skin	.26	**.41**	*.32*	.22	**.42**	*.33*
eyes	.04	**.53**	*.29*	−.14	**.84**	*.59*
nose	−.09	**.70**	*.45*	.01	**.73**	*.55*
mouth	−.11	**.76**	*.52*	.18	**.63**	*.56*
lips	−.02	**.78**	*.60*	.27	**.55**	*.56*
neck	**.34**	**.46**	*.45*	.15	**.61**	*.50*
chest/breast	−.04	**.47**	*.20*	**.72**	.10	*.61*
arms	**.54**	.23	*.45*	**.37**	**.33**	*.39*
hands	**.40**	**.39**	*.44*	.13	**.59**	*.45*
abdomen	**.86**	−.15	*.66*	**.92**	−.14	*.71*
waist	**.86**	−.08	*.69*	**.91**	−.11	*.72*
genitals	**.41**	**.52**	*.62*	**.54**	.15	*.42*
buttocks	**.82**	.02	*.68*	**.63**	.14	*.53*
hips	**.89**	−.04	*.76*	**.67**	.22	*.67*
thighs	**.90**	−.08	*.75*	**.78**	−.01	*.60*
legs	**.80**	.06	.69	**.69**	.07	*.55*
feet	**.43**	**.36**	*.43*	.16	**.52**	*.40*
Factor correlations	.41			.60		
Cronbach's α (length)*	.92 (7)	.84 (11)		.90 (9)	.86 9)	
Total Cronbach's α	.91			.93		

Bold: factor loadings ≥.30; italics: communalities (h^2).
*Cronbach's α value of each subscale based on items with factor loadings underlined.

Table 1 shows the rotated factor loadings for the pattern matrices. The 2-factor structure was almost the same for girls and boys, with two exceptions: satisfaction with chest/breast and genitals. Therefore, factor 1 was labeled «torso» and factor 2 was labeled «head and limbs». Cronbach's alpha values were satisfactory (α≥.84).

Total body satisfaction correlated highly with general physical appearance (*r* .74; *p*<.001). Total body problems correlated moderately and negatively with total body satisfaction (*r*=−.42; *p*<.001) and general physical appearance (*r*=−.40; *p*<.001). Internal consistency for body problem scores was satisfactory both for girls and boys (KR20=.74).

Test-retest Reliability

The 1-month and 7-month test-retest reliability was high (Table 2, left).

Table 2. Test-retest reliability and correlations of QÜIC with eating attitudes (EAT-40), influences of aesthetic body ideal (CIMEC), and shape concern (EDE-Q-SC) scores

		Test-retest reliability[1]		**Relation to external measures[2]**		
		1 month (*n*=190)	**7 months (*n*=195)**	**EAT-40**	**CIMEC**	**EDE-Q-SC**
Total body satisfaction		.79**	.70**	−.23**	−.38**	−.42**
Satisfaction with torso	girls	.85**	.76**	−.39**	−.51**	−.52**
	boys	.85**	.76**	−.25**	−.48**	−.53**
Satisfaction with head and limbs	girls	.77**	.68**	−.03	−.12	−.19
	boys	.66**	.61**	−.06	−.22*	−.21
Total body problems		.67**	.67**	.45**	.55**	.53**
General physical appearance		.71**	.63**	−.28**	−.44**	−.45**
Conformity with						
Height		*.66***	*.54***	*−.33***	*−.24***	*−.14*
Weight		*.80***	.73**	−.42**	−.51**	−.51**

Normal font: Pearson's correlations for quantitative measures; italics: Kappa[1] or point-biserial correlations[2] for categorical measures (yes *vs*. no).
*p<.05; **p<.001.

Convergent and Discriminant Validity

Acceptable convergent and discriminant validity with external measures was obtained (Table 2, right).

Relation to Sex and Age

Girls showed less body satisfaction (6.92 *vs*. 7.31; *p*=.007; 95% CI [0.11; 0.66]), less general physical appearance (6.61 *vs*. 7.15; *p*=.003; 95% CI [0.19; 0.90]), and more body problems (2.23 *vs*. 1.34; *p*<.001; 95% CI 95% [0.44; 1.34])

than boys (t-tests comparisons). More than half of the girls wished to weigh less, whereas only a third of the boys did (53.9% *vs*. 34.7%; $p<.001$). No differences were found for conformity with height (54.2% in girls *vs*. 48.5% in boys; $p=.388$), regarding sex (chi-square tests) (descriptive statistics are available upon request).

Age only correlated with total body satisfaction, satisfaction with head and limbs in girls, and general physical appearance, but the magnitudes were low ($r\leq.18$).

Discussion

The QÜIC satisfaction items presented an adequate 2-factor structure, with two components that can be broadly summarized as «torso» and «head and limbs». Only two noticeable differences emerged between girls and boys: satisfaction with chest/breast and with genitals loaded higher on factor 1 (torso) in boys and on factor 2 (head and limbs) in girls. Thus, for boys the factor «torso» included the same areas covered by the BASS (Cash, 1997) as upper, mid, and lower torso, in addition to satisfaction with genitals. Considering that chest size is a more important aspect of body image and self-esteem for men than satisfaction with breast size is for women (Tantleff-Dunn & Thompson, 2000) and that females tend to be more neutral than males toward their genitalia (Morrison, Bearden, Ellis, & Harriman, 2005), it seems that both items loaded according to its «importance» for body image in each group of respondents.

Beyond the use of an anatomic criterion to label the above underlying dimensions following Cash (1997), we can consider the possibility of control over one's appearance: items loading on factor 1 would mainly refer to body parts that are disguise/conceal with clothing textile in Western cultures, while modifiable with dieting, compared to the items loading on factor 2 that are not.

Total body satisfaction and body problems were moderately and inversely correlated, supporting the existence of the affective and cognitive dimensions (Cash et al., 2004), whereas general physical appearance was closer to the first one.

Satisfactory convergent and discriminant validity was achieved. Results regarding sex and age were aligned with previous research (Lawler & Nixon, 2011). Although ours is not a random sample, the percentages of conformity with weight and height were also similar to previous findings (Ricciardelli & McCabe, 2001).

To conclude, the Spanish version of the QÜIC appears to be an easily administrated and brief self-report questionnaire with satisfactory psychometric properties in our community sample. It includes the evaluation of both affective and cognitive components of body image, by means of a female or male figure, specially adapted for children and adolescents, who are population at risk of ED. The QÜIC makes it possible to assess satisfaction and concern for the overall body and for different parts of it, which may be an advantage over other current tools (Penelo, Espinoza, Portell, & Raich, in press).

Author's note

This work was partly supported by grants from the *Ministerio de Ciencia y Tecnología* (CBS02002-03689) and the *Ministerio de Educación y Ciencia* (SEJ2005-07099).

References

Allen, K. L., Byrne, S.M., McLean, N.J. & Davis, E.A. (2008). Overconcern with weight and shape is not the same as body dissatisfaction: Evidence from a prospective study of pre-adolescent boys and girls. *Body Image, 5*, 261-270. doi:10.1016/j.bodyim.2008.03.005

APA (American Psychiatric Association) (2000). *DSM-IV-TR: Diagnostic and statistical manual of mental disorders (4th ed. Revised)*. Washington, DC: Author.

Cash, T.F. (1997). *The body image workbook. An 8-step program for learning to like your looks*. Oakland, CA: New Harbinger Publications.

Cash, T.F., Melnyk, S.E. & Hrabosky, J.I. (2004). The assessment of body image investment: An extensive revision of the Appearance Schemas Inventory. *International Journal of Eating Disorders, 35*, 305-316. doi:10.1002/eat.10264

Castro, J., Toro, J., Salamero, M. & Guimerá, E. (1991). The Eating Attitudes Test: Validation of the Spanish version. *Evaluación Psicológica, 7*, 175-190.

Espinoza, P., Penelo, E. & Raich, R.M. (2010). Disordered eating behaviors and body image in a longitudinal pilot study of adolescent girls: What happens 2 years later? *Body Image, 7*, 70-73. doi:10.1016/j.bodyim.2009.09.002

Fairburn, C.G. & Beglin, S.J. (1994). Assessment of eating disorders: Interview or self-report questionnaire? *International Journal of Eating Disorders, 16*, 363-370. doi:10.1002/1098-108X(199412)16:4<363::AID-EAT2260160405>3.0.CO;2-#

Garner, D. & Garfinkel, P. (1979). The eating attitudes test: Validation of the symptoms of anorexia nervosa. *Psychological Medicine, 9*, 273-279. doi:10.1017/S0033291700030762

Giovannelli, T.S., Cash, T.F., Henson, J.M. & Engle, E.K. (2008). The measurement of body-image dissatisfaction–satisfaction: Is rating importance important? *Body Image, 5*, 216-223. doi:10.1016/j.bodyim.2008.01.001

Holmqvist, K. & Frisén, A. (2010). Body dissatisfaction across cultures: Findings and research problems. *European Eating Disorders Review, 18*, 133-146. doi:10.1002/erv.965

Lawler, M. & Nixon, E. (2011). Body dissatisfaction among adolescent boys and girls: The effects of body mass, peer appearance culture and internalization of

appearance ideals. *Journal of Youth and Adolescence, 40*, 59-71. doi:10.1007/s10964-009-9500-2

Miró, M.C. (2006). *Avaluació de la imatge corporal per a la detecció precoç de trastorns alimentaris*. Unpublished Doctoral Thesis, Universitat Autònoma de Barcelona.

Morrison, T.G., Bearden, A., Ellis, S.R. & Harriman, R. (2005). Correlates of genital perceptions among postsecondary students. *Electronic Journal of Human Sexuality, 8*. Available at http://www.ejhs.org/volume8/GenitalPerceptions.htm

Neumark-Sztainer, D., Paxton, S.J., Hannan, P.J., Haines, J. & Story, M. (2006). Does body satisfaction matter? Five-year longitudinal associations between body satisfaction and health behaviors in adolescent females and males. *Journal of Adolescent Health, 39*, 244-251. doi:10.1016/j.jadohealth.2005.12.001

Penelo, E., Espinoza, P., Portell, M. & Raich, R.M. (in press). Assessment of body image: Psychometric properties of the Body Image Questionnaire. *Journal of Health Psychology*. doi:10.1177/1359105311417913

Penelo, E., Villarroel, A.M., Portell, M. & Raich, R.M. (in press). Eating Disorder Examination Questionnaire (EDE-Q): An initial trial in Spanish male undergraduates. *European Journal of Psychological Assessment*. doi:10.1027/1015-5759/a000093

Ricciardelli, L.A. & McCabe, M.P. (2001). Dietary restraint and negative affect as mediators of body dissatisfaction and bulimic behavior in adolescent girls and boys. *Behaviour Research and Therapy, 39*, 1317-1328. doi:10.1016/S0005-7967(00)00097-8

Stanford, J.N. & McCabe, M.P. (2002). Body image ideal among males and females: Sociocultural influences and focus on different body parts. *Journal of Health Psychology,* 7, 675-684. doi:10.1177/1359105302007006871

Tantleff-Dunn, S. & Thompson, J.K. (2000). Breast and chest size satisfaction: Relation to overall body image and self-esteem. *Eating Disorders, 8*, 241-246. doi:10.1080/10640260008251231

Thompson, J.K. & van den Berg, P. (2002). Measuring body image attitudes among adolescents and adults. In T. F. Cash & T. Pruzinsky (Eds.), *Body image: A handbook of theory, research, and clinical practice* (pp. 142-154). New York: Guilford Press.

Tiggemann, M., Martins, Y. & Churchett, L. (2008). Beyond muscles: Unexplored parts of men's body image. *Journal of Health Psychology, 13*, 1163-1172. doi:10.1177/1359105308095971

Toro, J., Castro, J., Gila, A. & Pombo, C. (2005). Assessment of sociocultural influences on the body shape model in adolescent males with anorexia nervosa. *European Eating Disorders Review, 13*, 351-359. doi:10.1002/erv.650

Toro, J., Salamero, M. & Martínez, E. (1994). Assessment of sociocultural influences on the aesthetic body shape model in anorexia nervosa. *Acta Psychiatrica Scandinavica, 89*, 147-151. doi:10.1111/j.1600-0447.1994.tb08084.x

Villarroel, A.M., Penelo, E., Portell, M. & Raich, R.M. (2011). Screening for eating disorders in undergraduate women: Norms and validity of the Spanish version of the Eating Disorder Examination Questionnaire (EDE-Q). *Journal of Psychopathology and Behavioral Assessment, 33*, 121-128. doi:10.1007/s10862-009-9177-6

EVALUACION DE UN INSTRUMENTO DE VALORACION DE ASPECTOS IMPLICADOS EN EL EMBARAZO JUVENIL

Raquel A. Sarapura, Mª Pilar Jara Jiménez, Francisco Herrero Machancoses, Jacinto Pallarés Mestre y Ana Alarcón Aguilar
Universidad Jaume I de Castellón
Correo electrónico: reachel_27@hotmail.com

Resumen

Partiendo de investigaciones realizadas en Argentina y complementada con una metodología de trabajo cuantitativa, mediante un estudio piloto, nuestro objetivo es generar un instrumento que permita ampliar el estudio del embarazo en la juventud y la adolescencia, identificando los diversos aspectos desde el contenido teórico. Siguiendo los aspectos derivados del estudio teórico en relación a las secuelas en la interrupción del embarazo, detección de trastornos emocionales, embarazo deseado y no deseado, actitud de los padres, aspectos psicológicos del embarazo, la atención de la salud en el embarazo, maternidad en la adolescencia, funciones personales en el embarazo y reducción de daños en el embarazo juvenil y adolescente. Se genera un cuestionario que permite la observación de tales aspectos, focalizados en una serie de ítems que serán evaluados cuantitativamente con el objetivo de corroborar inferencias de primer nivel. Como indican, Buela-Casal, Sierra, Carretero-Dios y De los Santos-Roig, (2002), en España, lo normal suele ser la «importación» de instrumentos de evaluación. Desde esta perspectiva, la mayoría de los investigadores están más ocupados en adaptar que en crear, a pesar de las reiteradas advertencias sobre la falta de funcionalidad de muchas de estas adaptaciones, o sobre la ausencia de relevancia cultural que de ellas puede derivarse (Pelechano, 1997, 2002).

El objetivo general que planteamos es diseñar un cuestionario multifactorial cuyos ítems favorezcan el conocimiento en torno al embarazo precoz: casuística y características bio-psico-sociales. Para dar respuesta a los objetivos trazados mostraremos las fases para la confección del cuestionario así como su estudio psicométrico y estructural.

El trabajo fue realizado para dar inicio a un profundo y óptimo conocimiento de las madres adolescentes embarazadas en Castellón de la Plana**.** Desde un punto de vista transcultural, mediante el abordaje práctico llevado a cabo en el país de Argentina y utilizando de base la teoría del «Programa de vida» propuesta por Gabriel Castellá (2009), se propone el análisis y estudio de casos, en futuras investigaciones, implementadas en Castellón de la Plana. Tras el objetivo de promover

las tendencias prospectivas de las madres adolescentes embarazadas, conociendo y describiendo los modos existenciales de concepción más frecuentes.

En el futuro, se pretende favorecer el desarrollo de estrategias adecuadas de intervención familiar en el trabajo con adolescentes posibilitando la asunción de un rol positivo en las madres adolescentes embarazadas, y en consecuencia, un óptimo desarrollo de sus hijos. Así pues, mediante las intervenciones con las futuras constituciones familiares se busca favorecer una adecuada relación entre padres e hijos, soporte social de la sociedad del futuro, desde una perspectiva preventiva, orientadora y asistencial.

Las constituciones familiares conforman un terreno a explorar inacabable conforme va desarrollándose el ser humano, siendo una de las fuentes principales de socialización. Constituyendo, además, uno de los ámbitos más analizados a causa de las vertiginosas transformaciones sufridas a lo largo de la historia, que comienzan a adquirir mayor notoriedad, y conjuntamente a ello, sus implicaciones y consecuencias.

Desde el enfoque central de los contextos familiares para el desarrollo del ser humano, se toma como punto pivote el paso por la adolescencia que implica un periodo difícil de elaboración y duelo. Sumándole a ello la posibilidad de embarazos precoces a lo cual se superpone una afectividad aun más desconocida e incierta: la afectividad materna que ocupa un papel principal como mediadora de los vínculos, desde un plano amoroso y de contención.

La asunción de la maternidad en dicha etapa posee una doble condicionalidad; por un lado se caracteriza por ser un proceso de elaboración de duelos y conformación de una identidad. Por otro lado, requiere la elaboración de identidad materna asumiendo las implicaciones correspondientes a dicha etapa, diferenciándose de una madre adulta.

Frente a dos escenarios sociales, culturales y económicamente diferentes, numerosos estudios realizados en el ámbito dan muestra de los crecientes casos de embarazos adolescentes y la urgencia de atender a las necesidades psicológicas que surgen consecuentemente.

Estados Unidos es el país con el índice más elevado de embarazos, según la ONU unas 15 millones de adolescentes de entre 15 y 19 años dan a luz cada año en todo el mundo, y 4,4 millones se someten a un aborto.

En Argentina nos encontramos con un 70% de casos en los que el embarazo adolescente no es deseado, de entre los cuales el 15% de los nacimientos anuales tiene lugar entre niñas y adolescentes de entre 10 y 19 años. Escárraga (2004) señala: *«Alcanzando esta maternidad precoz un total de 700.000 casos anuales en todo el país y a la vez con más de 500.000 abortos realizados, dando lugar, consecuentemente a ello, el ascenso en el porcentaje de complicaciones en los abortos que pasaron de ser 53.900 en el año 1995 a distinguirse por una diferencia abismal en el año 2010.*

No obstante, las estadísticas realizadas en España por el Ministerio de Sanidad, reflejan al menos 18.000 casos de embarazos de adolescentes menores de 19 años, de entre las cuales solo el 25% (de 15 y 17 años) reconoce tener relaciones sexuales y el 12, 7% no utiliza ningún método anticonceptivo».

Velasco (2010) señala que estudios realizados en clínicas privadas del País Vasco, se encontraron con nueve de cada diez casos de interrupciones voluntarias en la gestación, (47% de adolescentes vascas de 15 a 19 años), que, crece aceleradamente anualmente. Hace una década apenas rozaba el 20%, actualmente al País Vasco está en la media del resto de España, seguido por Cataluña o Aragón (60%), Madrid (12,8%), Baleares (12.5%) y Murcia (11,7%).

Desde estos porcentajes, se infiere la necesidad y urgencia de implementar tratamientos eficaces que atiendan los casos de maternidad precoz. Para lo cual resulta imprescindible considerar la edad de inicio de la actividad sexual, acarreando en la mayoría de las ocasiones, un desconocimiento acerca del uso adecuado de métodos anticonceptivos.

Es preciso atender y responder a la demanda de los adolescentes, y la que respecta a los padres. Pues, el grupo familiar originario es el primer contacto de socialización y educación de las personas.

A partir de los datos transversales recabados en la investigación realizada en Argentina hace dos años, se diseñó un cuestionario para hombres y mujeres que nos permitía aproximarnos a la visión del embarazo precoz para la población española. Los datos iniciales se obtuvieron a través del trabajo de campo realizado con jóvenes de entre 14 y 19 años de Salta Capital que concurrieron por control obstétrico al Nuevo Hospital del Milagro.

Del conocimiento teórico derivado, así como el hecho de abundar en el conocimiento de la casuística del embarazo adolescente, se nos planteó la configuración de un instrumento que permitiese conocer las particularidades del embarazo en jóvenes y adolescentes.

Para llevar a cabo nuestro trabajo, como indican Buela-Casal, Sierra, Carretero-Dios y De los Santos-Roig, 2002; en España, lo normal suele ser la «importación» de instrumentos de evaluación. Desde esta perspectiva, la mayoría de los investigadores están más ocupados en *adaptar* que en *crear*, a pesar de las reiteradas advertencias sobre la falta de funcionalidad de muchas de estas adaptaciones, o sobre la ausencia de relevancia cultural que de ellas puede derivarse (Pelechano, 1997, 2002).

Así pues, nuestro objetivo se basó en diseñar un cuestionario multifactorial cuyos ítems favorezcan el conocimiento en torno al embarazo precoz: casuística y características bio-psico-sociales.

Producto de ello, y focalizándonos en una serie de ítems que fueron evaluados cuantitativamente tras el objetivo de corroborar inferencias de primer nivel, se desprendieron los siguientes objetivos específicos:

- Proponer y valorar ítems que, desde aspectos teóricos, recojan atributos relacionados con el embarazo precoz.
- Identificar qué puede contribuir, orientar e incentivar una posible intervención en búsqueda de una optima asunción y desempeño de la función maternal en la, en el caso que se continúe con el embarazo.
- Identificar qué puede contribuir, orientar e incentivar a las jóvenes y/o adolescentes para evitar embarazos no deseados.

Metodología

Debido a la necesidad de conocer el estado de opinión de personas adolescentes, o muy próximas a ella, se consideró la posibilidad de seleccionar una muestra conformada 42 estudiantes de 1º año de psicología de la Universidad Jaume I en Castellón de la Plana del año 2010.

En primera instancia la redacción de la gran cantidad de ítems, nos llevó, desde un grupo de debate, a una primera criba de ítems.

Con objeto de definir adecuadamente el contenido de la escala, y considerar su validez futura, los ítems se agruparon considerando los factores:

- Características de personalidad de la adolescente embarazada.
- Motivos del embarazo.
- Proyección de futuro.
- Perspectiva de la función materna.
- Educación sexual.

Posteriormente, se consideró la opinión de distintos jueces, quienes, mediante una variante de la técnica del grupo nominal, valoraron cada ítem con una escala del 1 (menor importancia) a 10 (mayor importancia).

Tras la aplicación del cuestionario, se obtuvo el valor medio de importancia de cada ítem, valores que más tarde permitieron ordenar las puntuaciones y mediante la aplicación «Explore» de SPSS (versión 19), se obtuvo el valor de las bisagras de Tukey que permitieron identificar aproximadamente los 25% ítems mejor valorados por el colectivo de estudiantes.

Finalmente se realizaron los diferentes análisis de datos. Para el estudio confirmatorio se utilizó el programa EQS (versión 6.1).

Resultados

La encuesta estuvo conformada por los ítems que obtuvieron puntuaciones iguales o superiores a 8.26. Un total de18 ítems que contienen todos los epígrafes

que apuntamos al inicio. A estos hemos añadido el ítem me siento feliz, como variable predictora en futuros trabajos de aplicación de la escala a embarazadas precoces.

El estudio psicométrico nos informa de que la escala alcanza un índice de fiabilidad α=0.947.

Los resultados obtenidos, en cuanto a posibles diferencias en la importancia que le otorgan los jóvenes a cada ítem, se encontró que existen diferencias significativas en:

– «Me siento a gusto con mi embarazo» $Z_{U\text{-}Mann}$ =-2,535 (sig=.018), cuando consideramos la variable sexo, de manera que ellas (Med=23,11) dan más importancia a este ítem que ellos (Med=9,60)

– «Me siento feliz» ZU-Mann =-1,996 (sig=.046), cuando compramos si practican o no alguna religión, de manera que los practicantes de alguna religión (Med=31,60) dan más importancia a este ítem que los que no practican (Med=20,14)

– «Obtuve la información sobre prevención de riesgo de enfermedades de trasmisión sexual a través de entre» $Z_{U\text{-}Mann}$ =-2,407 (sig=.016), cuando consideramos el centro en que finalizó sus estudios, de manera que los que estudiaron en un centro concertado (Med=30,75) dan más importancia a este ítem que los que estudiaron en un centro público (Med=19,74).

El estudio confirmatorio de los ítems (ver figura1) que refleja, por un lado, los cuatro factores considerados: características de personalidad de la adolescente embarazada (factor 1), proyección de futuro en el embarazo adolescente (factor 2), educación sexual (factor 3), perspectiva de la función materna (factor 4) (Ver el cuestionario que aparece al final del trabajo). Y por otro lado, los índices de ajuste del modelo puesto a prueba (ver tabla1).

Tabla 1. Índices de Ajuste

Satorra-Bentler scaled Chi-Square = 114.3036 Degrees of Freedom: 113
Probability value for the chi-Square statistic is .44800
FIT INDICES
Bentler-Bonett Normed Fit Index =.768
Bentler-Bonett Non Normed Fit Index = .996
Comparative Fit Index (CFI) = .996
Bollen (IFI) Fit Index = .997
Mcdonald (MFI) Fit Index =.985
Root Mean-Square Error Of Approximation (RMSEA) =.017
90% Confidence Interval Of RMSEA (.000, .080)

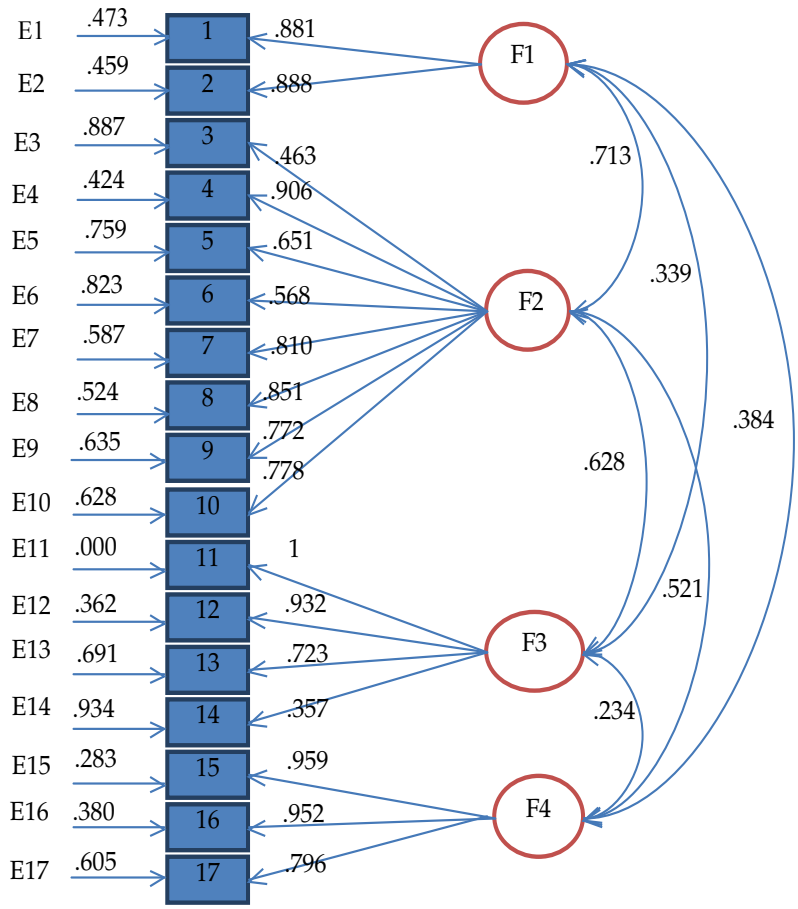


Figura 1. Estructural Factorial del cuestionario: Análisis Confirmatorio.

DISCUSIÓN

En cuanto la importancia que le otorgan los jóvenes a cada ítem, se encontró que existen diferencias significativas en cuanto al hecho de que las mujeres dan diferencialmente más importancia al hecho de sentirse a gusto con el embarazo. Del mismo modo que los practicantes de alguna religión diferencialmente otorgan más importancia al hecho de sentirse feliz. Es importante resaltar que los jóvenes que estudiaron en un centro concertado dan más importancia al hecho de obtener información sobre prevención de riesgo de enfermedades de trasmisión sexual.

El cuestionario generado es fiable y tiene validez factorial para cuyos ítems favorecen el conocimiento en torno al embarazo precoz, su casuística y las características bio-psico-sociales.

REFERENCIAS

Buela Casal, G., Sierra, J. C., Carretero Dios, H. y De los Santos Roig, M. (2002). Situación actual de la evaluación psicológica en lengua castellana. *Papeles del Psicólogo*, 27-33.

Carretero, H. y Pérez, C. (2005). Desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology ISSN*. (5 Vols.) (3). Granada, España. Asociación Española de Psicología Conductual

Castella, G. J. (2006*). La concepción y el sentido de la existencia: Teoría del Programa de Vida I*. Buenos Aires, Argentina: San Pablo.

EL PAIS (2007). *La mitad de las adolescentes embarazadas aborta*. Consultado el 19 de julio del 2010 en*:* http://www.clinicasabortos.com/ noticias_detalle.asp?id=33

Escarraga, T. (2004). *Embarazos en adolescentes.* Consultado el 19 de Julio del 2010 en http://www.paginadigital.com/articulos/2004/2004cuart/tecnologia3/tec1-11pD.asp

Escarraga, T. (2004). *Embarazos en adolescentes. Sexo desde muy jóvenes*. Consultado el 20 de Julio de 2010 en: http://www.consumer.es/web/es/salud/prevencion/2004/10/25/110838.php

Liendro, S. N. (2006). Los embarazos repetidos en las adolescentes. *Tesis de Licenciatura* en Psicología, Salta Capital, Universidad Católica de Salta, Argentina.

Pelechano, V. (1997). Prólogo. En G. Buela-Casal y J.C. Sierra (dirs.), *Manual de evaluación psicológica. Fundamentos, técnicas y aplicaciones* (pp. 31-35). Madrid: Siglo XXI.

Pelechano, V. (2002). Valoración de la actividad científica en psicología? Pseudoproblema,sociologismo o ideologismo? *Analisis y Modificacion de Conducta*, 28, 323-362.

Sarapura, R. A. (2008). Modos existenciales de concepción más frecuentes en las jóvenes madres. *Tesis de Licenciatura* en Psicología, Salta Capital, Universidad Católica de Salta, Argentina.

Velasco, C. (2010). *Embarazo no deseado, datos en España*. Consultado el 20 de julio de 2010 en http://pequelia.es/39970/embarazo-no-deseado-datos-en-espana/

ANEXO

Embarazo Precoz: Cuestionario General Multifactorial Definitivo

Indique su acuerdo o desacuerdo con las siguientes afirmaciones:

Item	Totalmente en desacuerdo	En desacuerdo	De acuerdo	Absolutamente de acuerdo
Cuando supe que estaba embarazada pensé en abortar.				
Mi pareja y yo planeamos el embarazo.				
A pesar de embarazo voy a seguir estudiando.				
Preferiría no estar embarazada.				
Sé lo que voy a hacer con mi futuro.				
Para mí ser madre es ser responsable.				
Para mí ser madre es educar.				
Para mí ser madre es atender las necesidades del bebé.				
Me considero capaz de cumplir con mi función como madre.				
Conozco los medios preventivos para el embarazo.				
Conozco los medios preventivos para las enfermedades de trasmisión sexual.				
Estoy informada acerca de las enfermedades por transmisión sexual.				

Obtuve la información sobre prevención del embarazo a través de:				
Padres	Amigos/as	Televisión	Internet	Otros

Conozco los siguientes métodos de prevención del embarazo.				
Preservativos	Pastillas Anticonceptivas	Dispositivo intrauterino	La pastilla del día después	Anillo vaginal

Este es un embarazo deseado			
Si, lo busqué	Si, aunque no lo busqué	No, aunque no puse medios para evitarlo	No, porque puse medios para evitarlo

VALIDEZ DE LA ADAPTACIÓN ESPAÑOLA DEL ABQ: UN ENFOQUE MULTI-RASGO/MULTI-MÉTODO

Gloria Seoane[1], Thomas Raedeke[2] María José Ferraces[1], Cristina de Francisco[1], Iria Arce y Constantino Arce[1]

[1] Universidad de Santiago de Compostela

[2] East Carolina University

Resumen

El presente estudio se ha realizado con el objetivo de investigar la validez convergente y discriminante de la versión española del ABQ, un instrumento elaborado por Raedeke y Smith para la medida del burnout en deportistas. Para ello, se ha utilizado una muestra española de 302 deportistas a los que se ha administrado el ABQ junto con otros dos cuestionarios, el MBI-GS y el DASS-21, que miden respectivamente burnout en entornos organizacionales y depresión, ansiedad y estrés. Para el análisis de datos se ha utilizado la aproximación multi-rasgo/multi-método que se ha abordado desde la perspectiva de los modelos de ecuaciones estructurales. Mediante el análisis comparativo de cuatro modelos, se ha poido ha podido concluir que la versión española del ABQ dispone de validez convergente con otras medidas similares y de validez discriminante entre rasgos y métodos.

El presente estudio se ha realizado con el objetivo de aportar pruebas sobre la validez convergente y discriminante de la adaptación española del ABQ mediante el análisis de relaciones entre sus rasgos y los rasgos medidos por otros dos cuestionarios: el MBI-GS y el DASS-21. El ABQ es un instrumento para la medida del burnout en deportistas (Raedeke y Smith, 2001) basado en la teoría tri-dimensional de Raedeke (1997): agotamiento físico y emocional (AFE), devaluación de la práctica deportiva (DPD) y reducida sensación de logro (RSL). El MBI-GS es un cuestionario ideado por Maslach, Jackson y Leiter (1996) para la medida del burnout en entornos laborales, que también posee tres rasgos: desgaste emocional (DE), cinismo (C) y eficacia profesional (EP); mientras que el DASS-21 ha sido desarrollado por Lovibond y Lovibond (1995) para medir depresión (D), ansiedad (A) y estrés (E).

Bajo hipótesis deberían observarse correlaciones más altas entre los rasgos teóricamente equivalentes del burnout medidos por el ABQ y el MBI-GS que entre los rasgos medidos por cada uno de los dos cuestionarios entre sí. La convergencia debería producirse entre AFE y AE, entre DPD y C y entre RSL y EP y del ABQ y MBI-GS, respectivamente. Y la discriminación debería producirse entre AFE,

DPD y RSL del ABQ y entre AE, C y EP del MBI-GS. Repecto a la relación entre ABQ y DASS-21 se esperarían relaciones directas entre los rasgos de ambos cuestionarios pero no tan elevadas como las correlaciones entre los rasgos equivalentes del ABQ y el MBI-GS.

MÉTODO

Participantes

302 deportistas (206 hombres y 96 mujeres) de diferentes modalidades y niveles deportivos, con edades que oscilaban entre los 14 y los 29 años (media = 19.06; desviación típica = 3.876).

Instrumentos

Versión española del Athlete Burnout Questionnaire (Arce, De Francisco, Andrade, Arce y Raedeke, 2010, De Francisco, 2010). El cuestionario consta de 15 ítems, divididos equitativamente en tres sub-escalas para la medida de las dimensiones del burnout propuestas por Raedeke (1997).

Versión española del MBI-GS (Salanova y Schaufeli, 2000; Moreno-Jiménez, Rodríguez-Carvajal y Escobar-Redonda, 2001). Se compone de 16 ítems para evaluar las tres dimensiones del burnout propuestas por Maslach y Jackson y Leiter (1996).

Versión española del DASS-21 (Bados, Solanas y Andrés, 2005). Está formada por 21 items repartidos equitativamente en tres sub-escalas: depresión, ansiedad y estrés.

Resultados

En primer lugar, se realizó un AFC donde se especificaron 9 factores (tres para cada cuestionario), 52 variables observadas (número total de items), y 52 términos de error. El modelo estaba sobreidentificado con 1378 elementos en la matriz de datos y 140 parámetros a estimar, que se distribuían de la siguiente manera: (1) 52 cargas factoriales, una para cada asociación factor-item hipotetizada, fijando las restantes cargas factoriales a cero, (2) 52 términos de error, uno para cada para cada variable observada (item), manteniendo incorrelacionados los términos de error entre sí, y (3) 36 correlaciones entre factores. Como método de estimación se utilizó ML implementado en AMOS (Byrne, 2010).

En la Tabla 1 se ofrece la matriz multi-rasgo/multi-método con las correlaciones entre los rasgos respectivos de los tres instrumentos de medida utilizados en la investigación y los coeficientes Alpha de Cronbach en la diagonal principal. De acuerdo con una de las hipótesis se observa como las correlaciones más altas se

producen entre AFE y DE (r_{xy}= .855, p<.01), DPD y C (r_{xy}= .755, p<.01), y RSL y EP (r_{xy}= -.644, p<.01),mismos rasgos medidos por distintos métodos (ABQ y MBI-GS, respectivamente), siendo mayores estas correlaciones que las observadas entre distintos rasgos medidos por el mismo método, que oscilan entre .257 y .553 en el caso del ABQ y entre -.126 y .545 en el caso del MBI-GS.

Tabla 1. Matriz multi-rasgo/multi-método

		ABQ			MBI-gs			DASS-21		
		AFE	**DPD**	**RSL**	**DE**	**C**	**EP**	**D**	**A**	**E**
ABQ	***AFE***	.841								
	DPD	.335	.707							
	RSL	.257	.553	.724						
MBI-gs	***DE***	.855	.335	.331	.824					
	C	.368	.775	.608	.545	.678				
	EP	-.064	-.314	-.644	-.126	-.303	.769			
DASS-21	***D***	.344	.338	.551	.372	.479	-.218	.783		
	A	.279	.129	.337	.304	.261	-.251	.630	.700	
	E	.533	.208	.453	.541	.397	-.161	.602	.552	.798

Respecto a la segunda hipótesis establecida, donde se especificaba la relación entre los rasgos del ABQ y DASS-21, se observa también que se confirma, dado que existen correlaciones directas situadas entre .129 y .551 que, aún siendo estadísticamente significativas, no llegan a ser tan elevadas como las observadas entre los rasgos equivalentes del ABQ y el MBI-GS, que oscilan entre -.644 y .855.

Finalmente, se procedió a realizar un análisis multi-rasgo/multi-método complementario desde la perspectiva de los modelos de ecuaciones estructurales, siguiendo las indicaciones de Widaman (1985). Se especificaron cuatro modelos. Un modelo base o Modelo 1, el menos restrictivo de los cuatro, denominado de rasgos correlacionados/métodos correlacionados (RC/MC), que asume que la varianza de cada ítem está determinada por tres componentes (factor rasgo, factor método y término error) y permite correlaciones libres entre los rasgos y entre los métodos, estando los rasgos y los métodos incorrelacionados entre sí. El segundo modelo (Modelo 2), denominado no rasgos/métodos correlacionados (NR/MC), más restrictivo y anidado al modelo base, asume la no existencia de rasgos y la correlación libre entre métodos. El Modelo 3, que se denomina rasgos perfectamente correlacionados/métodos correlacionados (RPC/MC), especifica una correlación perfecta entre rasgos y una correlación libre entre métodos. Por último, el Modelo 4, denominado rasgos correlacionados/métodos incorrelacionados (RC/MI), establece la hipóteis de que los rasgos están correlacionados, pero no los métodos.

En la Tabla 2 se ofrecen los índices del ajuste global de los cuatro modelos y en la Tabla 3, las diferencias en χ^2, con su significatividad estadística, las diferencias en los grados de libertad y las diferencias en CFI observadas entre el Modelo base o Modelo 1 y los restantes modelos.

Tabla 2. Índices de bondad de ajuste de los modelos multi-rasgo/multi-método

Modelos	χ^2	***gl***	χ^2/gl	***CFI***	***RMSEA***	***I.C.***
Modelo 1	2235.34	1189	1.88	.817	.054	.051,.058
Modelo 2	4383.19	1277	3.43	.458	.09	.08,.093
Modelo 3	2741.082	1217	2.252	.734	.065	.061,.068
Modelo 4	2223.47	1195	2.02	.786	.058	.055,.062

Tabla 3. Índices de bondad de ajuste diferencial entre los modelos multi-rasgo / multi-método

		Diferencia en	
Comparación de modelos	χ^2	***gl***	***CFI***
Prueba de validez convergente			
Modelo 1 vs Modelo 2	2147.75***	88	.35
Prueba de validez discriminate			
Modelo 1 vs Modelo 3	505.64***	60	.08
Modelo 1 vs Modelo 4	11.97	6	.028

*** p<.001.

La diferencia estadísticamente significativa entre el Modelo base o Modelo 1 y el Modelo 2 se interpreta como evidencia favorable a la validez convergente del ABQ; la diferencia estadísticamente significativa entre el Modelo base o Modelo 1 y el Modelo 3 como evidencia favorable a la validez discriminante entre rasgos y la diferencia estadísticamente no significativa entre el Modelo base o Modelo 1 y el Modelo 4 como evidencia a favor de la validez discriminante entre métodos.

CONCLUSIONES

A través de la presente investigación se ha podido obtener evidencia a favor de la validez convergente y discriminante de la versión española del ABQ, un cuestionario diseñado por Raedeke y Smith (2001) para la medida del burnout en deportistas. Los rasgos del ABQ convergen con los rasgos equivalentes del MBI-GS, que mide burnout en entornos laborales. A su vez, los rasgos del ABQ se diferencian entre sí (validez discriminante entre rasgos) y de otros rasgos teóricamente asociados, medidos por el DASS-21, como son la depresión, la ansiedad y el estrés (validez discriminante entre métodos).

NOTA DE LOS AUTORES

La presente investigación ha sido realizada con el apoyo del Ministerio de Ciencia e Innovación y del Fondo Europeo de Desarrollo Regional-FEDER (PSI2010-18807).

REFERENCIAS

Arce, C., De Francisco, C., Andrade, E., Arce, I. y Raedeke, T. (2010). Adaptación española del Athlete Burnout Questionnaire (ABQ) para la medida del burnout en futbolistas. *Psicothema, 22*(2), 250-253.

Bados, A., Solanas, A. y Andrés, R. (2005). Psychometric properties of the Spanish version of Depression, Anxiety and Stress Scales. *Psicothema, 17*, 679-683.

Byrne, B.M. (2010). *Structural equation modeling wit AMOS: baisc concepts, appliations , and programming.* Taylor and Francis Group, LLC.

De Francisco, C. (2010). *Adaptación psicométrica de una medida de burnout basada en el modelo ABQ de Raedeke y Smith*. (Tesis doctoral; Recurso electrónico). Santiago de Compostela: Servizo de Publicacións e Intercambio Científico, Universidade de Santiago de Compostela.

Lovibond, P.F. y Lovibond, S.H. (1995a). *Manual for the Depression Anxiety Stress Scales*. Sidney: Psychology Foundation of Australia.

Maslach, C., Jackson. S.E., y Leiter, M.P. (1996). *Maslach Burnout Inventory*. Third Edition. Palo Alto, CA: Consulting Psychologist Press.

Moreno-Jiménez, B, Rodríguez-Carvajal, R. y Escobar-Redonda, E. (2001). La evaluación del burnout profesional. Factorización del MBI-GS. Un análisis preliminar. *Ansiedad y Estrés, 7*(1), 69-78.

Raedeke, T. D. (1997). Is athlete burnout more than just stress? A sport commitment perspective. *Journal of Sport and Exercise Psychology, 19*(4), 396-417.Raedeke, T. D. y Smith, A. L. (2001). Development and preliminary validation of an athlete burnout measure. *Journal of Sport and Exercise Psychology, 23*(4), 281-306.

Raedeke, T. D. y Smith, A. L. (2001). Development and preliminary validation of an athlete burnout measure. *Journal of Sport and Exercise Psychology, 23*(4), 281-306.

Salanova, M. y Schaufeli, W.B. (2000). Exposure to information technologies and its relation to burnout. *Behavior & Information Technology*, 19, 385-392.

Widaman, K.F. (1985). Hierarchically tested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1-26.

CUESTIONES TEÓRICAS

LA CONCEPCIÓN DE LA MEDICIÓN PSICOLÓGICA

Juan Delgado[1], Joan Guàrdia[2] y Jordi Fauquet[3]
[1] Universidad de Salamanca
[2] Universitad de Barcelona
[3] Universitad Autónoma de Barcelona
Correo electrónico: jdelgado@usal.es

Resumen

Los procesos de medición en psicología siempre han estado presididos por una cierta debilidad generada por la precariedad de muchos de los supuestos bajo los cuales una estructura métrica cumple con los criterios matemáticos subyacentes. La tradición de Stevens y las posiciones derivadas del operacionalismo han mantenido un modo de afrontar esta cuestión que, en nuestra opinión, oscila entre medidas ausentes de toda condición métrica hasta medidas que se basan exclusivamente en el dominio del ajuste estadísticos para darles sentido. Estas y otras cuestiones asociadas se revisan en este trabajo con el objeto de ofrecer un panorama crítico a la medición psicológica.

La posición operacionalista de Stevens pretendía responder a la objeción de Campbell (1920, 1928), según la cual era imposible medir atributos psicológicos. En este sentido, el convencimiento de Campbell sobre tal imposibilidad, fundamentado en los trabajos originales de Helmholtz (1887) y Hölder (1901), se basaba en que la mayoría de «variables psicológicas» no son susceptibles de una operación natural de concatenación (ni tan sólo de bisección) y no poseen una estructura matemática aditiva del tipo que requieren los sistemas extensivos clásicos. Stevens utiliza las ideas del primer operacionalismo de Bridgman (1927) que presenta al «análisis operacional» como un filtro o método de control contra los errores que, de una manera u otra, habían conducido al colapso a la física de Newton. Resulta curioso comprobar que el impacto de la propuesta de Bridgman, que no fue el primero en enfatizar la importancia del análisis operacional (Eddington en 1920 había discutido nociones similares y Peirce en 1878 había avanzado posibles soluciones a problemas ontológicos), fue mucho mayor entre los psicólogos que entre los propios físicos a los cuales iba dirigida su propuesta. La posición de Bridgman apuntaba además otro gran lema bajo el que todos los psicólogos han estado expuestos «*el concepto es sinónimo de un conjunto de operaciones*». Sin embargo, Bridgman (1927) exigía que el conjunto de operaciones equivalente a un concepto fuese un conjunto *único*. Esta unicidad de procedimiento constituye una de las piedras angulares del denominado «operacionalismo metodológico» (Feest, 2005)

y es importante destacar que Bridgman nunca habló de definir conceptos por vía de las operaciones sino de análisis operacional o de método operacional, lo que no es lo mismo (al propio Bridman no le gustaba el término de operacionismo o operacionalismo).

A partir de trabajos previos (Stevens, 1935a, 1935b, 1936) Stevens escribe en 1939 dos trabajos de importancia fundamental. El primero de ellos, «The operational definition of psychological concepts» (Stevens, 1939a), representa una declaración de principios acerca de la constitución de una psicología basada en el operacionalismo, el positivismo lógico, el fisicalismo y el empirismo científico. En el segundo, «On the problem of scales for the measurement of psychological magnitudes» (Stevens, 1939b), apoyándose en su concepto operacionalista, revisa aspectos fundamentales de la medida y se pregunta sobre los posibles significados que pueden darse a los escalas propuestas para la medición de la sensación. A este respecto: *«An analysis of arithmetic from the point of view of syntactics and semantics reveals the following posible distinctions: 1. Numerals are certain signs we commonly make on paper. The names of the numerals are «one», «two», «three», etc. 2. Numerousness is a property or attribute which we are able to discriminate when we regard a collection of objects. 3. Numerosity is a property defined by certain operations performed upon groups of objects. Numerals, as signs, are related to numerosity and to numerousness according to certain semantical rules (...). The following is proposed as a possible way of looking at the syntactics and semantics of numerals.»* (Stevens, 1939b). De lo anterior se desprende que Stevens:

1. Evita el concepto de número, y define numeral como signo, aunque las relaciones entre cantidad y número se habían cuestionado y el concepto de «numeral» ya se había propuesto (Campbell, 1928).

2. Cuando habla de «*numerousness*» como «*una propiedad o atributo que somos capaces de discriminar cuando observamos un grupo de objetos*», utiliza el concepto «discriminar»; concepto central en su teoría de la Psicología como ciencia operacional (Stevens, 1935).

3. La descripción de *«numerosity*» como *«una propiedad definida por ciertas operaciones realizadas sobre ciertos grupos de objetos*», enfatiza el hecho de que es el «usuario» quien realiza/ejecuta esas operaciones.

Cuando atribuimos numerales, lo hacemos en función de *nuestra* capacidad de discriminación en un momento dado. Y, a su vez, esa capacidad de discriminación se ve favorecida o contrarrestada por las operaciones que realicemos. El famoso ejemplo de Stevens (1939b) ilustra lo anterior: si tomamos unos puñados de alubias y formamos con ellas varios montones, y podemos discriminar en qué grupo hay aparentemente más judías o en cuál menos, y asignamos numerales en este caso, lo hacemos sobre el concepto de «*numerousness*» (operación: ordenar). Pero si elegimos dos montones y emparejamos una judía de uno con otra de otro hasta que uno de los dos se agota, y en el otro quedan judías, éste se dice que tiene más «*nome-*

rousity». Si asignamos numerales a estos grupos en función de su «numerosidad», esa asignación se fundamenta en el recuento (operación: contar).

Una opción a considerar, aunque desconocida, es la teoría representacional de la medida (véase, por ejemplo, Krantz, Luce, Suppes y Tversy, 1971; Luce y Suppes, 2002; Luce, Krantz, Suppes y Tversky, 1990; Narens, 1985; Narens y Luce, 1986; Pfanzagl, 1968; Suppes y Zinnes, 1963; Suppes, P., Krantz, D.H., Luce, R.D., & Tversky, A. 1989) que, entre otras aportaciones, ha formalizado los conceptos de escala de medida de Stevens, la parte más brillante y permanente de su teoría. Todo este trabajo nos lleva a dos conclusiones aparentemente obvias: medir es algo distinto de lo que quiso Stevens y no es lo mismo la cantidad que la cualidad. Para medir, lo primero que se necesita es un atributo cuantitativo. Michell (2001) lo expresó de forma clara: «*…, la medición se aplica sólo a atributos cuantitativos*.». Ante el argumento de que las variables psicológicas no pueden medirse como las físicas, se puede responder que es cierto, pero que junto con la medición extensiva, existe la medición implícita, intensiva, formalizada y definida por primera vez por Luce y Tukey (1964) denominada medida conjunta, que intenta solucionar simultáneamente el problema de la medida y el de la composición construyendo escalas de medida para las variables relevantes de modo que se satisface el principio de composición (Tversky, 1967).

¿Por qué no se sigue el camino antes delineado? Porque rompe con hábitos muy bien establecidos. La forma correcta de hacer las cosas comienza con una decisión: Cuando pretendamos crear una medida de algo o alguien, ¿comenzamos por la operación concreta de la construcción de la medida, como propone Stevens (y como hacen muchos psicólogos)? ¿O comenzamos por el principio, es decir, por el aspecto o atributo específico para el que se trata de identificar diferentes magnitudes? La solución positivista y operacionalista fue un sí a la primera opción. Pero esa respuesta es profundamente errónea: no se puede comenzar por la medición sin determinar qué y cómo es, qué propiedades o características posee lo que se pretende medir.

Y esto implica conocer, o definir, o proponer una red de significado acerca de ese atributo o aspecto (un concepto en última instancia) que deseamos medir: lo que significa y sus relaciones con el resto de conceptos igualmente significativos con los que se relacionará. La posición realista en la ciencia supone que ese concepto alude a una realidad independiente de quien pretende medirlo. La definición científica de conceptos nunca puede ser subjetiva: es inter-subjetiva, y se basa en su significado normativo, es decir, en que no dependa de quién los defina (como propone el subjetivismo operacionalista de Stevens) sino de qué sea lo definido.

La posición operacionalista reduce todo a obtener un número, y ya Lord propuso que los números no saben de dónde vienen. Y puede ser cierto, pero nosotros sí debemos saberlo, para no operar sin sentido con números. Cuando los sumamos, como se suman los ítems de una escala, usamos las operaciones de la concatenación aditiva, sin siquiera antes considerar si la variable que pretendemos medir es capaz de sostener tales operaciones. Se tienen claramente identificados los errores

lógicos que se cometen al hacer esto. Las técnicas estadísticas (sobre todo las multivariadas más utilizadas: análisis de regresión, análisis factorial, componentes principales, modelos de ecuaciones estructurales) aplican operaciones matemáticas que se fundamentan en las propiedades de las variables con estructura ordinal y aditiva. El problema es que las conclusiones obtenidas sólo se pueden aplicar si las variables poseen esas propiedades. Con atributos cualitativos no tienen sentido ciertas operaciones matemáticas, lo sabemos desde que aprendimos las escalas de Stevens. Las conclusiones se fundamentan en la estructura de la variable, y si ésta no la tiene, ¡las conclusiones no se sostienen!

En este problema están implicadas más cosas que las resolubles con un cambio de metodología estadística, matemática o programática. Específicamente, está implicado un cambio de hábitos de pensamiento, de planteamiento general de la investigación en las ciencias del comportamiento. No basta con aplicar el modelo de Rasch (por poner un ejemplo evidente) a los valores de una variable «definida» según los criterios de Stevens. Un ejemplo de los riesgos asociados con este error lo proporcionó hace tiempo Wood (1978). En este trabajo ajustó con un modelo de Rasch datos aleatorios, como los obtenidos al tirar una moneda al aire. Y el modelo ajustó, creando así una variable latente de intervalos iguales de habilidad de «tirar una moneda». Lo interesante del ejemplo es que pone de manifiesto que el fallo no está en la metodología, sino en las condiciones de *significado* bajo las que se inicia el escalamiento. Éste es un caso más de una construcción de medida que literalmente no tiene sentido.

A estas alturas, a la pregunta de si podemos realmente producir escalas de medida de intervalos iguales con ciertas propiedades como medida, para variables que signifiquen algo, en las ciencias del comportamiento, debemos responder con un sí definitivo. Pero con el costo de comenzar por el principio, por la definición de los conceptos, objetos o atributos que pretendemos medir. Abandonando el operacionalismo y el positivismo. No vale con decir que el positivismo ha sido superado: superémoslo con hechos.

Construyamos escalas de medida considerando primero *qué* es lo que intentamos medir. Esto evitará con toda seguridad que sigamos en la situación actual, en un maremagno de escalas alternativas virtualmente arbitrarias para un mismo supuesto objeto a medir y que no dispongamos de medios coherentes de elegir una u otra. Y evitará una situación en la que no se puede continuar, que ninguna de esas escalas consiga una medición precisa (de *qué*, por otra parte), excepto por azar.

La diferencia entre las ciencias del comportamiento y las llamadas ciencias «duras» (la física como ejemplo ubicuo) no está sólo en los hábitos de medición, que no tienen por qué ser tan diferentes. Estriba en sus extremadamente diferentes fundamentos conceptuales. Mientras esas ciencias físicas utilizan conceptos *técnicos* como cimientos para la construcción de sus cuerpos de conocimiento, en gran parte de las áreas de las ciencias del comportamiento el endeble y arenoso suelo sostiene una red de *conceptos seudo-psicológicos ordinarios* de gramáticas notablemente complicadas, condiciones de significado mal definidas, no compartidas

inter-subjetivamente, muchas veces de escasa referencia empírica y, peor aún, que en demasiados casos no son más que artificios seudo-matemáticos o seudo-conceptuales.

Así, en nuestra opinión, y como conclusión general, la tarea de la medición se debería articular claramente:

- En primer lugar, siempre se ha de definir un significado normativo, técnico, para el concepto.
- Debe ser susceptible de ser medido con precisión. Se podrá así construir una medida normativa apropiada para este concepto.
- Medida que está sujeta a evaluación y comprobación empírica para que se pueda confiar en que alude exactamente al significado normativo del concepto definido.

Sólo así, mediante una clara teoría de la medición, se puede mantener la medida creada, mientras sea defendible, porque ya se sabe, la ciencia y el conocimiento no suelen detenerse.

REFERENCIAS

Bridgman, P.W. (1927). *The logic of modern physics.* New York: Macmillan.

Campbell, N.R. (1920). *Physics: the elements.* London: Cambridge University Press.

Campbell, N.R. (1928). *An account of the principles of measurements and calculations.* London: Longsmans Green.

Debreu, G. (1960). Topological methods in cardinal utility theory. En Arrow, K.J., Karlin, S., & Suppes, P. (Eds.): *Mathematical Methods in the Social Sciences.* Stanford: Stanford University Press.

Eddington, A. (1920). *Space, time and gravitation.* Cambridge: Cambridge University Press.

Feest, U. (2005). Operationism in Psychology:what the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences, 41,* 131-149.

Fisher, R, I. (1892). *Mathematical investigations in the theory of value and prices.* Trans. Conn. Acad. Sci. 9, 1-24.

Fisher, I. (1927). A statistical method for measuring «marginal utility» and testing the justice of a progressive tax. En Hollander, J.H. (Ed.): *Economic essays contributed in honour of John Bates Clark*. New York: McMillan.

Green, C.D. (1992). Of immortal mythological beasts: operationism in psychology. *Theory & Psychology, 2,* 291-320.

Helmholtz, H.v. (1887). Zählen und messen erkentnnis-theoretisch betrachet. *Philosophise Aufsätze Eduard Zeller Gewidmet.* (Traducido al inglés por BRYAN, L.C. (1930). *Counting and measuring.* New York: Van Nostrand Reinhold Company).

Hölder, O. (1901). Die axiome der quantität und dir lehre van mass (The axioms of quantity and the theory of mass). *Sächsische Akademie Wissenschaften* zu Leipzig. *Mathematisch-Psysische Klasse 53*, 1-64.

Krantz, D.H. (1964). Conjoint measurement: the Luce-Tukey axiomatization and some extensions. *Journal of Mathematical Psychology 1*, 248-277.

Krantz, D.H., & Tversky, A. (1971). An exchange on functional and conjoint measurement. *Psychological Review 78*, 457-458.

Krantz, D.H.; Luce, R.D; Suppes, P. & Tversky, A. (1971). *Foundations of Measurement, Vol. I: Additive and polynomial representations*. New York: Academic Press.

Langfeld, H.S. (1945). Introduction to symposium on operationism. *Psychological Review, 52,* 241-248.

Luce, R.D. (1966). Two extensions of conjoint measurement. *Journal of Mathematical Psychology 3,* 348-370.

Luce, R.D., & Suppes, P. (2002). Representational measurement theory. In H. Pashler & J. Wixted (Eds.), Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology (3ª ed., pp. 1–41). New York: Wiley.

Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology, 1,* 1-27.

Luce, R.D., Krantz, D.H., Suppes, P., & Tversky, A. (1990). *Foundations of Measurement, Vol. III: Representation, Axiomatization, and Invariance.* New York: Academic Press.

Michell, J. (2001). *Teaching and misteaching measurement in psychology*. New York: Wiley & Sons.

Narens, L. (1985). *Abstract measurement theory.* Cambridge: MIT Press.

Narens, L., & Luce, R.D. (1986). Measurement: the theory of numerical assignements. *Psychological Bulletin 99*, 166-180.

Peirce, C.S. (1955). How to make our ideas clear. En J. Buchler (Ed.), *Philosophical writings of Peirce* (pp. 23-41). New York: Dover. (Original 1878)

Pfanzagl , J. (1971). *Theory of measurement*. 2ª Ed. (1ª de 1968) Wurzburg: Physica – Verlag.

Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology 1*, 233-247.

Stevens, S.S. (1935a). The operational basis of psychology. *American Journal of Psychology, 47,* 323-330.

Stevens, S.S. (1935b). The operational definition of psychological concepts. *Psychological Review, 42,* 517-527.

Stevens, S.S. (1936). Psychology: the propaedeutic science. *Philosophy of science, 3,* 90-103.

Stevens, S.S. (1939a). Psychology and the science of science. *Psychological Bulletin, 36,* 221-263.

Stevens, S.S. (1939b). On the problem of scales for the measurement of psychological magnitudes. *Fifth International Congress for the Unity of Science,* Cambridge: Massachussets, USA. September, 3-9.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (pp. 1-76). New York: Wiley.

Suppes, P., Krantz, D.H., Luce, R.D., & Tversky, A. (1989). *Foundations of Measurement, Vol. II: Geometrical, Threshold, and Probabilistic Representations.* New York: Academic Press

Tversky, A. (1967). A general theory of polinomial conjoint measurement. *Journal of Mathematical Psychology 4*, 1-20.

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology*, Vol 31(1), May 1978, 27-32.

SUPRESIÓN, SUPRESIÓN CLÁSICA Y MEDIACIÓN: COMPARACIÓN DE PERSPECTIVAS DE ANÁLISIS

Sergio Murgui Pérez y María C. Fuentes
Universitad de Valencia
Correo electrónico: Sergio.Murgui@uv.es

Resumen

El llamado efecto supresor, en los últimos años y con el incremento de las técnicas SEM, no ha sido estudiado en demasiada profundidad, si lo comparamos con el llamado efecto mediador. No obstante, en ambos casos se parte de correlaciones entre al menos tres variables, siendo una variable la dependiente, otra la independiente y una tercera variable la que causa el correspondiente efecto. El objetivo del presente trabajo es analizar mediante un marco común ambos fenómenos. Para ello, se han analizado los supuestos del análisis de supresión y se comparan con los establecidos para el análisis de mediación. El procedimiento de análisis supone grupos de tres variables correlacionadas entre sí, bien correspondientes a una mediación, bien correspondientes a la supresión. Así, se ha aplicado la regla de supresión y se ha calculado el tamaño del efecto, el efecto indirecto (*a*b*) y el valor de *c'*. Los resultados obtenidos muestran que ambos fenómenos son muy similares y que pueden diferir únicamente por el valor y el signo de las correlaciones utilizadas. Finalmente, se comentan cómo a la mediación se le proporciona una apariencia de causalidad, lo que no sucede con la supresión.

Una vez establecida una relación entre dos variables, X e Y (coeficiente de correlación o de regresión, *c*) a menudo resulta interesante investigar el efecto de otra variable que, relacionada con ambas (valores *a* y *b*), sea pronosticada por X y predictora de Y. Dicha tercera variable, denominada Z, puede tener efectos de mediación o de supresión. Aunque existen diversas definiciones de variable supresora (Maaser y Baker, 2001), en general, se asume que una variable es supresora si, al incluirla en la ecuación de regresión entre la variable independiente X y una variable dependiente Y, incrementa el coeficiente de regresión de X sobre Y (Conger, 1974). Por otra parte, cuando la relación entre las tres variables corresponde a una situación de redundancia, es posible analizar la mediación: la introducción de la variable Z reduce la relación previa entre X e Y (Baron y Kenny, 1986). Por tanto, el efecto de la tercera variable se puede establecer a partir de la comparación entre r_{YX} y β_{YX} (entre el *c* previo y el *c'* posterior).

A pesar de la similitud entre ambos fenómenos, son poco frecuentes los trabajos que estudian la supresión en psicología (Cheung y Lau, 2008). De hecho, una vez asumido que $r_{YX} > 0$, el efecto de supresión presenta tres tipos diferentes (Cohen, Cohen, West y Aiken, 2003): *clásica o tradicional*, que ocurre cuando $r_{YZ} = 0$; *reciproca o cooperativa*, cuando r_{YZ} (o bien r_{ZX}) es negativa y las otras dos correlaciones son positivas; y la supresión *negativa o neta*, cuando $\beta_{YX.Z}$ y/o $\beta_{YZ.X}$ resultan negativas. Además, la mediación se suele asociar a algún tipo de cadena causal por los investigadores al describir sus hipótesis y sus resultados (Wood, Goodman, Backman y Cook, 2008), en cambio, pocos autores (Ato y Vallejo, 2011) mencionan la supresión como parte del razonamiento causal.

En el estudio de la supresión desde el enfoque de mediación, MacKinnon, Krull y Lockwood (2000) plantean ambas dentro de un mismo esquema, con dos condiciones previas: relación significativa r_{YZ} y relación significativa r_{XZ}. Por tanto, descartan el primero de los pasos de Baron y Kenny (1986). De hecho, existe bibliografía en la que se analiza la significación estadística del efecto indirecto (por ejemplo, Bollen y Stine, 1990; Hayes y Preacher, 2010; Sobel, 1982) llegando a utilizarse como sinónimo de efecto mediado o mediación, recuérdese que

$$c = c' + a * b \tag{1}$$

Shrout y Bolger (2002) afirman que es posible encontrar valores negativos en *c'*, y dichos valores quedarían explicados por el hecho de que en la población el intervalo de *c'* incluye el cero (como en la supresión clásica). Además, la comparación de *c* y *c'* parece equivalente a la de $|r_{YX}|$ y $|\beta_{YX.Z}|$ planteada desde la supresión (Friedman y Wall, 2005; Tzelgov y Stern, 1978).

En la supresión se distinguen dos enfoques predominantes: el de Velicer (1978) y el de Conger (1974) si bien a partir de la similitud de ambos, Shieh (2006) ha planteado como supresión una situación tal que

$$r_{YX}^2 < \frac{(r_{YX} - r_{YZ}r_{ZX})^2}{(1 - r_{ZX}^2)^2} \tag{2}$$

Por otra parte, la interpretación de la supresión puede ser bastante sencilla aunque lo es menos si se utilizan términos «causales» (Paulhus, Robins, Trzesniewski y Tracy, 2004), en especial si de trata de la denominada supresión clásica. Además, la redundancia, la mediación y la supresión neta parten de la misma situación previa: el valor positivo de todas las relaciones entre las variables.

A partir de lo expuesto, nuestro objetivo es establecer los paralelismos entre la supresión clásica, neta y cooperativa y la mediación y su interpretación «causal».

MÉTODO

El análisis se basa en plantear distintos valores de correlación para cada una de las tres relaciones entre las variables X, Y, Z. Se presentan ternas de correlaciones todas ellas positivas, situación usual en la mediación. A continuación, se ha esta-

blecido una magnitud de 0,30 para r_{YX}, de 0,60 para r_{YZ} y de 0,45 y 0,55 para r_{XZ}. En todos los casos, se ha alternado el signo de la relación (situación 5 a 16). Finalmente, se han mantenido constante la relación r_{YX}, igualándola a cero, y se han alternado los valores de r_{YZ} y de r_{XZ} (situación 17 a 24), aplicándose la regla propuesta por Shieh (2006). Se ha calculado el valor de *c'* y el de *a*b* y el tamaño del efecto para la mediación ($R^2_{4.5}$) de MacKinnon (2008), mediante:

$$r^2_{YZ} - (R^2_{\text{Y.XZ}} - r^2_{YX}) \tag{3}$$

RESULTADOS

Los resultados obtenidos se muestran en la Tabla 1. En primer lugar podemos apreciar que las tres correlaciones positivas de la Situación 1 a 4 cumplen con la regla establecida para la supresión (en negrita). Los valores de las correlaciones, por otra parte, son similares a los valores de una situación de mediación (compárese los valores de la Situación 4 con los valores de la Situación 5 o 6). Además, el efecto indirecto es positivo tanto en las Situaciones 1 a 4 (de supresión) como en Situación 5 y 6 (de mediación).

En las Situaciones 5 a 18 se aprecia las consecuencias del cambio de signo en las correlaciones. Así, una situación de mediación (Situación 5 y 6) pasa a ser supresión al invertirse el signo de una de las relaciones, bien entre X y Z (Situación 7 y 8), de r_{XZ} (Situación 9 y 10) o de r_{YX} (Situación 11 y 12).

Si lo que se invierte es el sentido de la variable (lo que provoca el cambio de signo de dos de las correlaciones) bien de Z (Situaciones 13 y 14), de X (Situaciones 17 y 18) o de Y (Situaciones 15 y 16), el efecto de mediación se mantiene. Además, el efecto indirecto puede ser el mismo, tanto por su magnitud como por su signo, en una situación de mediación (Situación 5 y 6) como de supresión (Situación 19 y 20), que presenta signo y valor paralelos al efecto *a*b*.

En cuanto al tamaño del efecto propuesto por MacKinnon (2008), su valor se mantiene al margen del cambio del signo de las correlaciones, siempre que se trate de una mediación (Situaciones 5, 6 y Situaciones 13 a 18). En cambio, llega a presentar valores negativos cualquier tipo de supresión que no sea neta, es decir, basadas en alguna correlación negativa o cero (Situaciones 19 a 26).

Finalmente, en el caso de que no exista relación entre X e Y, nos encontramos en todos los casos con una situación de supresión clásica (Situaciones 19 a 26). Respecto al efecto indirecto, éste puede presentar valores positivos o negativos, según el signo de las correlaciones. En cuanto al valor de *c'* puede igualmente variar su signo, si bien presenta un valor de signo contrario al del efecto indirecto. En esta situación, el cálculo del tamaño del efecto proporciona valores negativos.

Tabla 1. Efecto de Z según la variación de las correlaciones entre X, Y, Z

Situación	X-Z (*a*)	Z-Y (*b*)	X-Y (*c*)	$\beta_{YX.Z}$ (*c'*)	**Regla de supresión**	*a*b*	$R^{2\ a}_{4.5}$	**Efecto de Z**
1	0,75	0,60	0,30	-0,343	**0,090<0,118**	0,450	0,039	Supresión neta
2	0,75	0,25	0,35	0,371	**0,123<0,138**	0,188	0,062	Supresión neta
3	0,85	0,55	0,35	-0,423	**0,123<0,179**	0,468	0,073	Supresión neta
4	0,60	0,25	0,45	0,469	**0,203<0,220**	0,150	0,062	Supresión neta
5	0,45	0,60	0,30	0,038	0,090<0,001	0,270	0,089	Mediación
6	0,55	0,60	0,30	-0,043	0,090<0,002	0,330	0,089	Mediación
7	-0,45	0,60	0,30	0,715	**0,090<0,511**	-0,270	-0,317	Supresión reciproca
8	-0,55	0,60	0,30	0,903	**0,090<0,816**	-0,330	-0,479	Supresión reciproca
9	0,45	-0,60	0,30	-0,715	**0,090<0,511**	-0,270	-0,317	Supresión reciproca
10	0,55	-0,60	0,30	-0,903	**0,090<0,816**	-0,330	-0,479	Supresión reciproca
11	0,45	0,60	-0,30	-0,715	**0,090<0,511**	0,270	-0,317	Supresión reciproca
12	0,55	0,60	-0,30	-0,903	**0,090<0,816**	0,330	-0,479	Supresión reciproca
13	-0,45	-0,60	0,30	0,038	0,090<0,001	0,270	0,089	Mediación
14	-0,55	-0,60	0,30	-0,043	0,090<0,002	0,330	0,089	Mediación
15	0,45	-0,60	-0,30	-0,038	0,090<0,001	-0,270	0,089	Mediación
16	0,55	-0,60	-0,30	0,043	0,090<0,002	-0,330	0,089	Mediación
17	-0,45	0,60	-0,30	-0,038	0,090<0,001	-0,270	0,089	Mediación
18	-0,55	0,60	-0,30	-0,043	0,090<0,002	-0,330	0,089	Mediación
19	-0,45	-0,60	-0,30	-0,715	**0,090<0,511**	0,270	-0,317	Supresión neta
20	-0,55	-0,60	-0,30	-0,903	**0,090<0,816**	0,330	-0,479	Supresión neta
19	0,45	0,60	0,00	-0,339	**0,000<0,115**	0,270	-0,091	Supresión clásica
20	0,55	0,60	0,00	-0,473	**0,000<0,224**	0,330	-0,156	Supresión clásica

DISCUSIÓN

Los resultados obtenidos muestran cuán tenue puede ser la diferencia entre un efecto de mediación y de supresión y, por otra parte, el riesgo evidente de descartar ciertas situaciones porque no se ajustan a un patrón «causal», entendido como mediación estadística. Ahora bien, si se entiende que una mediación «causal» como la cadena de relaciones que vincula las tres variables ¿por qué no puede incluirse la supresión como un tipo de mediación (causal)?

De hecho, una de las condiciones que Baron y Kenny (1986) establecen para la mediación «causal» es la existencia previa de una relación entre la variable X y la variable Y, solo así es posible encontrar una variable Z que medie en la relación anterior. Pero si este requisito se omite (Shrout y Bolger, 2002) lejos de simplificar el análisis lo vuelve más complejo. Así, Mathieu, DeShon y Bergh (2008), citando a Hyman (1955), se preguntan qué número de variables deben insertarse y estudiarse entre la variable X y la variable Y, y lo que es más relevante, en qué momento la interpretación de dicha cadena está completa.

Creemos que ha quedado demostrada la estrecha relación entre el efecto de mediación y el de supresión, de lo que se deriva la importancia de analizar ambos fenómenos de forma conjunta. Si se intenta extrapolar del primero una «causa» nada impide que se haga lo mismo del segundo, y en idénticos términos, pues el cambio de signo o de la magnitud de una correlación no debería afectar a la lógica causal. Finalmente, en torno a la mediación «causal» se puede plantear la siguiente pregunta: si ninguna técnica estadística establece una causalidad entre las variables ¿por qué el análisis de mediación debería ser distinto?

REFERENCIAS

Ato, M. y Vallejo, G. (2011). Los efectos de terceras variables en la investigación psicológica. *Anales de Psicología, 27* (2).

Baron, R.M., y Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bollen, K.A., y Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology, 20*, 115–140.

Cheung, G.W. y Lau, R.S. (2008). Testing mediation and suppression effects of latent variables: Bootstrapping With structural equation models. *Organizational Research Methods, 11*, 296-325.

Cohen, J., Cohen, P., West, S.G. y Aiken, L.S. (2003). *Applied multiple regression correlation analysis for the behavioral sciences*. 3rd Ed. Mahwah: Lawrence Erlbaum.

Conger, A.J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34*, 35-46.

Fiedman, L. y Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician, 59* (2), 127-136.

Hayes, A.F. and Preacher, Kristopher J.(2010) 'Quantifying and Testing Indirect Effects in Simple Mediation Models When the Constituent Paths Are Nonlinear', *Multivariate Behavioral Research, 45* (4), 627-660.

Hyman, H. (1955). *Survey design and analysis: Principles, cases and procedures.* Encino, CA: Glencoe.

Kasser, T. y Ryan, R.M. (1993). A dark side of the American dream: Correlatos of financial success as a central life aspiration. *Journal of Personality and Social Psychology, 65*, 410-422.

Maassen, G.H. y Bakker, A.B. (2001). Suppressor Variables in Path Models: Definitions and interpretations. *Sociological Methods & Research*, *30,* 241-270.

MacKinnon, D.P. (2008). *Introduction to statistical mediation analysis.* Mahwah, NJ: Erlbaum

MacKinnon, D.P., Krull, J.L. y Lockwood, C.M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1,* 173-181.

Mathieu, J.E., DeShon, R.P. y Bergh, D.D. (2008). Mediational Inferences in Organizational Research : Then, Now, and Beyond. *Organizational Research Methods, 11* (2), 203-223.

Paulhus, D.L., Robins, R.W., Trzesniewski, K.H. y Tracy, J.L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research, 39*, 303-328.

Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and Psychological Measurement, 66* (3), 435-447.

Shrout, P.E. y Bolger, N. (2002). Mediation in Experimental and nonexperimental Studies: new procedures and recommendations. *Psychological Methods, 7*, 422–445.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13*, 290–313.

Tzelgov, J. y Stern, I. (1978). Relationship between variables in three variable linear regression and the concept of suppressor. *Educational and Psychological Measurement, 38*, 953-958.

Velicer, W.F. (1978). Suppressor variables and the semipartial correlation coefficient. *Educational and Psychological Measurement, 38*, 953-958.

Wood, R.E., Goodman, J.S., Beckman, N. y Cook, A. (2008). Mediation testing in management research. A review and proposals. *Organizational Research Methods, 11*, 270-295.

MÉTODOS ESTADÍSTICOS

NON PARAMETRIC THREE WAY ANALYSIS OF VARIANCE WITH REPEATED MEASURES

Juan C. Oliver-Rodríguez
Universidad Jaume I de Castellón
Email: oliverr@uji.es

Abstract

Research problems that require a non-parametric analysis of multifactor designs arise frequently in the behavioral sciences. There is however a lack of available procedures in commonly used statistical packages. The present study proposes a generalization of the aligned ranked interaction test for the analysis of the sources of variation typically obtained in a three way *Anova* with repeated measures. Its statistical properties were tested by simulation methods by combining normal, exponential (asymmetric) or double exponential (symmetric and heavy tailed) distributions with presence and absence of sphericity in the covariance matrix and sample sizes of 10 and 30 participants per group. Classical and aligned ranked versions for univariate *F*, adjusted *F* by Lecoutre procedure and multivariate Hotelling statistics were compared. For normal distributions the classical statistics showed improved performance in most cases. For nonnormal distributions and large sample sizes the aligned rank statistics showed power improvements from 5 to 48% while having similar levels of Type I Error. For nonnormal distribution and small sample sizes an increase in Type I error rate for the aligned rank tests was however observed. The latter results are interpreted in the context of alternative procedures in the selection or development of small sample methods.

Problems in need of a nonparametric analysis of the interaction arise frequently in basic and applied behavior research. One proposed method consists of performing an *F* parametric contrast on the ranked observations after the main effect has been subtracted (Salter & Fawcett 1993). This procedure has been termed the aligned rank test and has been shown to be robust in terms of Type I error rates and statistical power in nonnormal distributions (Beasley, 2002). It is also easily implemented in standard statistical packages.

The objectives of the present study is first proposing a generalization of this method for the analysis of the sources of variation typically obtained in a three way *Anova* with repeated measures. Second, its statistical properties are tested by means of simulation methods.

The sources of variation for a mixed design with two between subject fixed effect factors (A and B) and one within subject fixed effect factor (M) can be specified as follows:

$$\begin{aligned} y_{ijkl} = \mu_{....} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + s_{l.(ij)} \\ + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \\ + \varepsilon_{kl(ij)} \end{aligned} \quad (1)$$

for $i = 1, 2, \ldots, a$ levels of factor A

$j = 1,2,\ldots,b$ levels of factor B

$k = 1, 2, \ldots, m$ levels of the within subject factors M

and $l = 1, 2, \ldots, n$ participants per experimental condition

Additionally random effects are Identically and Independently Distributed (IID) as follows:

$$\left.\begin{array}{l} s_{l.(ij)} \sim IID(0, \sigma_s^2) \\ \varepsilon_{kl(ij)} \sim IID(0, \sigma_e^2) \end{array}\right\} \textit{and both are mutually independent}$$

In a nonparametric context null location hypotheses are most flexibly expressed as equalites of cumulative distribution functions. These are summary row or column averages of the ordered values obtained at the different levels or level combinations of a source of variation (Brunner, Domhof and Langer, 2002; Shah & Madden, 2004). They are listed in Table 1.

The procedure for performing the aligned rank transformations is analogous to the one used for testing the interaction in two way designs. Here it is generalized by creating a new aligned rank variable for each experimental effect. Each variable is obtained by ranking the observations after removing the sources of variability not contained in the expected mean squares for that effect. The transformations for the between subject and within subject effects are described in Table 1. A standard *F* test will then be applied to each of the aligned rank variables instead of to the original raw variable. The corresponding test statistics will be called *Aligned Rank F´s* (F_{AR}). Only when the assumption of identical distributions above hold which include equal shapes and dispersion matrices can rejection of a null hypothesis be interpreted as a mean increase or decrease in the variable of interest between experimental conditions (Fay & Proschan, 2010).

When the sphericity assumption does not hold both the univariate correction to the *F* test proposed by Lecoutre (1991) and the multivariate Hotelling statistics will be used. They will be called *L* and *H*, and their aligned rank versions L_{AR} and H_{AR} statistics. The latter have been shown to have good properties under sphericity violation when applied to aligned ranks in two way mixed model designs (Beasley, 2002).

SIMULATION PROCEDURES

A simulated experiment was conducted for each of 96 conditions defined by all possible combinations of presence or absence of main effects, two and three way interactions, sphericity in the covariance matrix, two sample sizes (10 and 30 participants) and three distributions: normal, exponential (asymmetric) and double exponential (heavy tailed symmetric). Two levels of each between subject factor and four levels of the within subject factors were used. A thousand replications per condition were run.

Presence or absence of effects were respectively defined by adding or subtracting a constant c = .125 ó c = 0 to or from two or more different levels of each source of variation. Random variability from each distribution type was generated with a mean of 0, a standard deviation of 1 and a covariance matrix with presence of sphericity (ε= 1) or its absence (ε = 0.67). The algorithm used was an extension of the Fleishman power method running on SAS IML software (Headrick & Sawilowski, 1999).

Data Analysis Plan

Comparisons were made between F vs F_{AR} for between subject tests and for within subject tests when the sphericity assumption held. Comparisons were made between L vs L_{AR}, and H vs H_{AR} for within subject tests under no sphericity conditions.

In each of the 48 simulation runs where an experimental effect was absent a binomial test was used to detect Type I Errors with $\alpha = .05$. In each of the 48 simulation runs were an experimental effects was present a McNemar test was used to test differences in power rates between raw and aligned rank statistics since the data fed to both was the same and their corresponding results were therefore correlated. The frequency of simulation runs with detection of Type I Errors or improved relative power of raw vs aligned rank statistics was then tabulated for each of the three main effects, each of the three two way interactions and for the three way interaction. Aggregate results for the three types of effects are listed in Table 2.

RESULTS

Large Sample (n = 30)

F vs F_{AR}. When the sphericity assumption held Type I Error frequencies were similar with normal distributions and similar or slightly better for the F_{AR} test with the exponential or double exponential distribution. In terms of power the F test had a slight advantage with normal distributions. The power advantage favored the F_{AR} test with the other two distributions and was especially large in the exponential case with averages ranging from 26 to 45% (Table 2).

L vs L_{AR}. When the sphericity assumption did not hold and the distribution was normal the *L* statistics performed slightly better in terms of Type I Error. No detectable differences were observed for the exponential and double exponential distributions. In terms of power the *L* test had a slight advantage with normal distributions. The advantage favored de L_{AR} with the other two distributions which was specially large and around 48% in the exponential case (Table 2).

It is also interesting to note that an improvement was observed in normal distributions and small samples for both the *L* and L_{AR} test over the raw and aligned ranked versions of the Huyhn-Feldt tests (*HF* and HF_{AR}) reported in standard statistical packages. An improvement was also observed for the L_{AR} test with the exponential distribution and small samples over the HF_{AR} test.

H vs H_{AR}. When the sphericity assumption did not hold and the distribution was normal the H_{AR} test showed similar or even improved performance in terms of Type I Error. No detectable differences were observed for the other two distributions. In terms of power the *H* test had a slight advantage with normal distributions. The power advantage favored the H_{AR} with the other two distributions which was specially large and around 28% for the exponential case (Table 2).

Small Sample (n=10)

F vs F_{AR}. When the sphericity assumption held and the distribution was normal *F* showed improved performance in terms of Type I Error. Increased values for the F_{AR} test were observed for the exponential distribution but no statistically detectable differences were found for the double exponential one. In terms of power the *F* test had a slight advantage with normal distributions. The power advantage favored the F_{AR} test with the other two distributions which was specially large for the exponential case with average values ranging from 35 to 43%. (Table 2)

L vs L_{AR}. When the sphericity assumption did not hold the L_{AR} statistic showed statistically detectable higher levels of Type I Error for the exponential and double exponential distributions, not for the normal one. In terms of power the L_{AR} test was similar to the *L* test for normal distributions and had an advantage for the other two distributions which was specially large for the exponential case with values around 40% (Table 2)

H and H_{AR} statistics. When the sphericity assumption did not hold the H_{AR} statistic had statistically detectable larger levels of Type I Error for the normal and double exponential distributions not for the exponential one. In terms of power the H_{AR} was similar to the *H* test for normal distributions and had an advantage for the other two distributions which was large for the exponential case with values around 34% (Table 2).

Table 1. Aligned rank transformations and null hypotheses for the sources of variation in a three way Anova with one within subject factor

Effect	Transformation	Hypotheses
Between Subject		
A	$R\left(y_{ijkl}^{A}\right) = Rank(\mu_{....} + \alpha_i + s_{l.(ij)})$	$F(y_{i..}) - F(y_{i'..}) = 0$ for all y and any i, i´ levels of factor A
B	$R\left(y_{ijkl}^{B}\right) = Rank\left(\mu_{....} + \beta_j + s_{l.(ij)}\right)$	$F(y_{.j.}) - F(y_{.j'.}) = 0$ for all y and any j, j´ levels of factor B
A × B	$F(y_{ij.}) - F(y_{i..}) - F(y_{.j.}) + F(y_{...}) = 0$ for all y and any i, j levels of factors A and B	$R\left(y_{ijkl}^{AB}\right) = Rank(\mu_{....} + \alpha\beta_{ij} + s_{l.(ij)})$
Within Subject		
M	$R\left(y_{ijkl}^{M}\right) = Rank\left(\mu_{....} + \gamma_k + \varepsilon_{kl(ij)}\right)$	$F(y_{..k}) - F(y_{..k'}) = 0$ for all y and any k, k´ *levels* of factor M
A × M	$F(y_{i.k}) - F(y_{i..}) - F(y_{..k}) + F(y_{...}) = 0$ for all y and any i, k levels of factors A and M	$R\left(y_{ijkl}^{AM}\right) = Rank\left(\mu_{....} + \alpha\gamma_{ik} + \varepsilon_{kl(ij)}\right)$
B × M	$R\left(y_{ijkl}^{BM}\right) = Rank\left(\mu_{....} + \beta\gamma_{jk} + \varepsilon_{kl(ij)}\right)$	$F(y_{.jk}) - F(y_{.j.}) - F(y_{..k}) + F(y_{...}) = 0$ for all y and any j, k levels of factors B and M
A × B × M	$R\left(y_{ijkl}^{ABM}\right) = Rank\left(\mu_{....} + \alpha\beta\gamma_{ijk} + \varepsilon_{kl(ij)}\right)$	$F(y_{ijk}) - F(y_{i..}) - F(y_{.j.}) - F(y_{..k}) + F(y_{ij.}) + F(y_{i.k}) + F(y_{.jk}) - F(y_{...}) = 0$ for all y and any i, j, k levels of factors A and B and M

Notes: Superscripts on the *y* variable for the aligned ranked transformations indicate that these are performed separately for each source of variation. Estimates for main, two way interaction and three way interaction effects will be used in the calculations with sample data.

Table 2. Frequency of Detection of Type I Error and of Differences in Power Rates for Raw Scale and Aligned Ranked Statisics according to Binomial and McNemar tests

					Power			
			Type I	Rates[a]	McNemar Test		Average Gain	
Statistic	Distribución	B	W		B	W	B	W
				n = 30				
F vs F_{RA}	Normal	1 – 1	2 – 2		10 – 0	12 – 0	-3.99%	-2.23%
	Exp.	0 – 0	1 – 0		0 – 12	0 – 16	45.83%	26.32%
	Double Exp.	0 – 0	1 – 0		0 – 12	0 – 16	8.90%	5.29%
L vs L_{AR}	Normal		0 – 1			5 – 0		- 2.86%
	Exp.		0 – 0			0 – 16		48.18%
	Double Exp.		0 – 0			0 – 16		10.53%
H vs H_{AR}	Normal		1 – 0			11 – 0		-3.57%
	Exp		0 - 0			0- 16		28.44%
	Double Exp.		0 – 0			0 – 14		5.29%
				n = 10				
F vs F_{RA}	Normal	1 – 2	1 – 2		0 – 0	2 – 0	–	-2.57%
	Exp	0 – 2	1 – 3		0 – 12	0 -16	35.03%	43.76%
	Double Exp.	2 – 2	0 – 0		0 – 8	0 – 11	6.86%	7.76%
L vs L_{AR}	Normal		2 – 1			0 – 0		–
	Exp.		1 – 2			0 – 16		40.66%
	Double Exp.		0 – 4			0 -10		9.47%
H vs H_{AR}	Normal		2 – 3			1 – 1		–
	Exp.		1 – 0			0 – 16		34.27%
	Double Exp.		1 – 2			0 – 7		8.28%

Notes. F and F_{AR} tests are reported only in conditions where assumptions hold. L , L_{AR}, H and H_{AR} are reported only in within subject tests in conditions of no covariance sphericity. B and W refer to between and within subject tests respectively. Each entry of the table summarizes results for main and interaction effects. Power gain refers to power differences divided by the level obtained with raw scale statistics. Exp. and Double Exp. stand for Exponential and Double exponential distributions.

DISCUSSION

For normal distributions the classical statistics showed improved performance over the nonparametric ones in most cases. Lecoutre correction reduced Type I Error rates in comparison with the Huyhn and Feldt adjusted F test reported in standard analyses packages supporting previous analytical claims (Lecoutre, 1991). Its use by the researcher may benefit accuracy of results in the analysis of three way repeated measures designs with small samples. After twenty years of realizing the mistake it would be desirable that statistical companies decide to correct it.

When the assumption of normality did not hold the aligned ranked statistics showed improved performance over the classical tests with large samples sizes

only. The advantage was specially large with the asymmetric exponential distribution with power gains higher than twenty five percent. For nonnormal distributions and small sample sizes however, the aligned rank statistics showed an increased number of type I errors. Use recommendations are therefore analogous to the comparison of Mann-Whitney or Wilcoxon to the t test in one factor designs with two levels. These rank tests have also been shown more efficient for moderate or large samples and asymmetric or heavy tailed distributions (Higgins, 2004).

For small samples alternative nonparametric methods are available to the researcher by using specialized statistical software (Brunner, Domhof & Langer, 2002; Crimin, Abebe & McKean, 2008; Erceg-Hurn & Mirosevich, 2008). An integrative consideration of these methods may however provide insights for testing modifications of the aligned ranked method to the analysis of small samples, unbalanced data not considered here and other experimental designs. Comparisons of these perspectives in future studies may be valuable in providing general use and easily implemented techniques for the applied researcher.

REFERENCES

Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research, 37* (2), 197-226.

Brunner, E., Domhof, S. & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments*. New York: Wiley.

Crimin, K, Abebe, A. & McKean, J. W. (2008). Robust general linear models and graphics via a user interface. *Journal of Modern Applied Statistical Methods, 7* (1), 318-330.

Erceg-Hurn, D. M & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist, 63* (7), 591-601.

Fay, M. P. & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys, 4*, 1-39.

Headrick, T. C. & Sawilowski, S. S. (1999). Simulating correlated multivariate nonnormal distributions: extending the Fleishman power method. *Psychometrika, 64*, 25-35.

Higgins, James J. (2004). *Introduction to modern nonparametric statistics*. Belmont, CA: Duxbury Press.

Lecoutre, B. (1991). A correction for the ε approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics, 16*, 371-372.

Salter, K. C. & Fawcett, R. F. (1993). The ART test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in statistics: Simulation and Computation, 22*, 137-153.

Shah, D. A. & Madden, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology, 94*(1), 33-43.

UN EJEMPLO DE LA UTILIDAD DEL MODELO DE REGRESIÓN LOGÍSTICA ORDINAL EN ESTUDIOS CON VARIABLES DE TIPO FRECUENCIAL ACUMULATIVO UTILIZANDO EL PROGRAMA SPSS

Jacinto Pallarés, Jesús Rosel, Mª Pilar Jara, Francisco Herrero, Maria José Calero

Universidad Jaume I de Castellón
Correo electrónico: pallares@uji.es

Resumen

La regresión logística es un modelo que ofrece una solución muy satisfactoria a los problemas multivariantes. Aquí, se aplica a una variable dependiente de tipo ordinal, número de reincidencias, y se considera la probabilidad de un suceso (reincidencia), y del resto de sucesos que lo preceden o lo siguen en la ordenación de las correspondientes categorías. Los datos se han extraído del protocolo de exploración utilizado por el Equipo Técnico del Juzgado de menores de Castellón y se ubican dentro de las fuentes documentales. La muestra es de 2370 menores de género masculino y 617 de género femenino, con edades comprendidas entre 12 y 21 años. Las variables con mayor poder predictivo de la reincidencia son impulsividad, consumo de sustancias y déficit en habilidades sociales, y como factores de protección la ansiedad y el cambio de provincia. Se proponen algunas recomendaciones de intervención con menores infractores.

Dentro de la teoría criminológica, la perspectiva más ajustada a la realidad del presente estudio es la que ofrece, por un lado, la *teoría del etiquetamiento* (Lemert, 1967) y, por otro, *la criminología del* desarrollo (Andrews y Bonta, 2006). El objetivo general del presente estudio es comprobar el grado de influencia de variables independientes de tipo socio-familiar, de personalidad y académico-laborales en el número de reincidencias de los menores infractores, con el fin de poder intervenir positivamente sobre las mismas.

MÉTODO

La muestra se compone de 2370 menores de género masculino y 617 de género femenino, con edades de 12 a 21 años, explorados por el equipo técnico del Juzgado de Menores de Castellón entre enero 1998 y febrero de 2007, utilizando el

instrumento denominado «Informe Social y Psicoeducativo del menor», cuyos apartados son:

1) Datos de filiación personal y familiar.
2) Area sociofamiliar.
3) Variables psicológicas y comportamentales.
4) Variables personales.

La variable dependiente «número de reincidencias» es de tipo frecuencial, configurando, estadísticamente, una escala categorial acumulativa de tipo ordinal. El menor se considera reincidente bajo la condición de haber recibido, al menos, dos medidas judiciales.

Tabla 1. Variables independientes

Variables individuales	Variables socio-familiares
Género.	Cambio en el lugar de filiación.
Edad de la primera infracción.	Menor de los hermanos.
Expediente de protección.	Fomentan la escolarización.
Variables de personalidad	Inestabilidad económica/laboral.
Influencia y/o dependencia del grupo de iguales.	Existencia de límites y normas.
Déficit de habilidades sociales.	Cumplimiento de los límites y normas por el menor.
Impulsividad (problemas de control de impulsos).	Emigración.
Tratamiento psicológico anterior (problemas emocionales).	Cambio de provincia.
Problemas depresivos.	Pareja de hecho del menor.
Problemas sensoperceptivos.	Paternidad del menor.
Baja autoestima.	Variables académicas y laborales
Ansiedad.	Fracaso escolar.
Variables de consumo de sustancias	Expulsiones/despidos.
Consumo de sustancias psicoactivas.	

Las variables independientes de la Tabla 1 se abordan individualmente, por grupos y en un modelo general, en concordancia con la hipótesis de que influyen en la reincidencia. El criterio para seleccionarlas consistió en incluir aquellas que habían demostrado poder predictivo en otros estudios y eliminar aquellas con muchos casos perdidos. En función de los resultados se planteará un modelo integrado que incluya aquéllas con mayor poder predictivo.

La regresión logística binaria puede modificarse para ser aplicada a una variable dependiente de tipo ordinal (mediante una regresión logística multinomial), considerando la probabilidad de un suceso, y del resto de sucesos que lo preceden, o lo siguen, en la ordenación de las correspondientes categorías.

El modelo de regresión logística binaria plantea que la variable de respuesta tiene dos niveles, 1 = ocurrencia de un determinado suceso, y 0 = no ocurrencia del

suceso; e intenta predecir la probabilidad de ocurrencia en función de una serie de variables predictoras. El modelo puede expresarse:

$$\ln[(Y')] = logit[\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1)$$

Ahora bien, en la regresión logística ordinal, en lugar de considerar la probabilidad de un suceso, se considera que si existen *j* niveles (frecuencias) de una variable dependiente de tipo ordinal, el modelo llevará a cabo *J-1* predicciones, cada una estimando las probabilidades acumuladas en un determinado nivel o por debajo de la categoría *j-ésima* de la variable dependiente. Este modelo se utiliza para estimar las «razones» (odds) de estar por encima o debajo de un determinado nivel de la variable de respuesta o dependiente.

Los valores de las categorías (en nuestro caso reincidencias) se ordenan en rangos, pero la distancia real entre las categorías resulta desconocida.

Es habitual encontrar estudios de regresión logística multinomial en los que las variables independientes de tipo categórico, se definen como variables «dummy» agrupándolas en 3 ó 4 categorías como máximo, debido a la dificultad de interpretar adecuadamente los coeficientes. Este escollo puede solventarse recurriendo a la regresión logística ordinal

A efectos didácticos se explicará la interpretación de los valores obtenidos con una sola variable independiente, en este caso *género*. El modelo logístico ordinal para una sola variable independiente (X_1), para *j* igual a 3 ó menos reincidencias, es:

$$ln(\theta_j) = \alpha_j \mid [\beta_p(X_p)] \Rightarrow \pi(j \leq 3) = \frac{1}{1 + e^{-(\alpha_j - \beta_j x_1)}} \quad (2)$$

Donde *j* va desde *1* al número de categorías (reincidencias) menos *1*; y *p* es el número de variables independientes (en este caso una sola).

Cada logit posee su propio intercepto o valor de umbral ('threshold value', α_j), pero idénticos coeficientes para las distintas variables (β_1). Esto significa que el efecto de la variable independiente es el mismo para las diferentes funciones logit. Este efecto común a cada una de las categorías da lugar al denominado frecuentemente «modelo de razones (odds) proporcionales». El coeficiente β determina la forma de las curvas de regresión. Para describir la asociación entre las variables se utiliza $e^{-\beta}$, que indica el incremento de las odds para las probabilidades acumuladas por unidad de incremento en la variable independiente *X*. Los componentes α_j, denominados valores de umbral, no dependen de los valores de la variable independiente para cada uno de los casos, y aunque equivalen al intercepto de la regresión lineal, cada uno posee su logit propio y se utilizan para la estimación de los valores predichos.

Tomando como ejemplo la V.I. genero, podemos observar en la Tabla 2 como el parámetro β_j (*-0 ,*779) describe el efecto de la V.I. (*x*) (género) sobre la V.D. (*y*) (nº de reincidencias). Si exponenciamos β_j, $e^{-\beta} = e^{-0,779} = 2,17$, significa que el valor

de la «odds ratio» a lo largo de todas las categorías de reincidencia es 2,17. Esto indica que la probabilidad de reincidencia en cada una de las categorías es 2,17 veces más probable en el caso de los menores infractores que de las menores, por término medio.

Tabla 2. Estimaciones de los parámetros de la ecuación de regresión ordinal de la V.D. Número de reincidencias en función de la V.I. Género

		Estimación	Error típ.	Wald	gl	Sig.	Intervalo de confianza 95%	
							Lím. inferior	Lím. superior
Umbral	[Nº_reincidencias = 0]	1,977	,063	990,218	1	,000	1,854	2,100
	[Nº_reincidencias = 1]	2,810	,086	1062,524	1	,000	2,641	2,979
	[Nº_reincidencias = 2]	3,624	,124	859,709	1	,000	3,382	3,866
	[Nº_reincidencias = 3]	4,016	,149	731,323	1	,000	3,725	4,307
	[Nº_reincidencias = 4]	4,375	,176	615,864	1	,000	4,029	4,721
	[Nº_reincidencias = 5]	4,784	,215	495,188	1	,000	4,363	5,206
	[Nº_reincidencias = 6]	5,314	,279	363,445	1	,000	4,767	5,860
	[Nº_reincidencias = 7]	5,801	,355	267,594	1	,000	5,106	6,496
	[Nº_reincidencias = 8]	6,089	,409	221,489	1	,000	5,287	6,891
	[Nº_reincidencias = 9]	6,496	,501	168,269	1	,000	5,514	7,477
	[Nº_reincidencias = 10]	7,190	,708	103,225	1	,000	5,803	8,577
	[Nº_reincidencias = 11]	7,883	1,000	62,096	1	,000	5,922	9,844
Ubicación	Género	-,779	,181	18,535	1	,000	-1,134	-,425

Función de vínculo: Logit.

El parámetro α_j o 'umbral' (1,977, 2,810, etc.) es propio para cada una de las categorías de reincidencia e indica el carácter acumulativo del modelo. Si queremos conocer p.e. la «odds» de que un menor infractor pueda reincidir más de 3 veces, sustituimos los valores. En primer lugar para el género masculino (codificado como *0*):

$$p(\leq 3) = \frac{1}{1 - e^{-(4{,}016)}} = ,9822 \qquad (3)$$

$$1 - ,9822 = ,0138$$

En segundo lugar para el género femenino (codificado como *1*):

$$p(\leq 3) = \frac{1}{1 - e^{-(4{,}016)-(-{,}779 \cdot 1)}} = \frac{1}{1 - e^{-4{,}795}} = ,9918 \qquad (4)$$

$$1 - ,9918 = ,0082$$

De cada 1.000 niñas, aproximadamente 992 reinciden menos de 3 veces, el resto (8 niñas) reincide 4 o más veces. Si calculamos la «odds ratio» entre niños y niñas:

$$OR\,(j \geq 4) = {}^{0{,}0138}/_{0{,}0082} = 1.7. \qquad (5)$$

Esto significa que los niños con 4 o más reincidencias resultan 1,7 veces más frecuentes que las niñas. Si hiciéramos el cálculo de las *OR* para todas las reincidencias, veríamos que oscilan alrededor de 2,17, que es el valor de $e^{-\beta}$ del modelo. De cada 1.000 niños que han entrado en el sistema judicial, siendo etiquetados como infractores, 982 reinciden tres o menos de tres veces, el resto (18) reinciden 4 o más veces. En definitiva las niñas reinciden muchísimo menos que los niños.

En el caso del módulo estadístico de SPSS-17, el modelo logit de regresión ordinal, implementa el modelo universal **politómico** (PLUM o Polytomous Universal Model) y utiliza un sub-modelo acumulativo propio en el cual el denominador es la proporción acumulada de todos los casos; también denominado «modelo de razones proporcionales» (proportional odds model), o «modelo de razones acumulativas» (cumulative odds model) (Agresti, 2002; Long & Freese, 2006), siendo su ecuación general:

$$logit\left[\pi \leq {}^{j}/_{x_1, x_2, \ldots, x_p}\right] = ln\left(\frac{\pi\left(Y \leq \frac{j}{x_1 x_2}, \ldots, x_p\right)}{\pi\left(\frac{Y}{x_1 x_2}, \ldots, x_p\right)}\right) = \alpha_j + \left(-\beta_1 x_1 - \beta_2 x_2 \ldots - \beta_p x_p\right) \quad (5)$$

Donde:

$\pi\left(Y \leq \frac{j}{x_1 x_2}, \ldots, x_p\right)$ *es la probabilidad de encontrarse dentro de o por debajo de la categoría j* (j *=1, 2, ...j-1*)*, dados un conjunto de predictores* (x_1, x_2, x_p)*:*

α_j son los puntos de corte, umbrales o niveles de respuesta;

$\beta_1, \beta_2 \ldots, \beta_3$ son los coeficientes logit de las correspondientes variables independientes.

De este modo, se predicen los logits acumulados a lo largo de las *j-1* categorías de respuesta. Con la transformación de los valores logits acumulados, se obtienen las razones (odds) acumuladas, así como las probabilidades acumuladas de pertenecer a una determinada categoría, o encontrarse por debajo de dicha categoría.

Con este procedimiento se llegó a un modelo integrado de predicción que incorpora las variables más significativas, tal y como podemos ver en la Tabla 3, con una ganancia altamente significativa respecto al modelo nulo (χ^2= *393,83, p = 0,000*).

De acuerdo con los valores de Tabla 3 se plantean los siguientes ejemplos de estimación:

Ejemplo 1:

Cálculo de la probabilidad de 3 o más reincidencias para un menor infractor (de género masculino, codificado como 0), que cometió su primera infracción a los 14 años (que codificado en meses da el valor de 168) y en quien se den todas las variables propuestas en el modelo que, en caso de ser dicotómicas, se codificarán como *0* en caso de ausencia o como *1* en caso de presencia (X1 = 0; X2 = 168; … ; X8 = 1). El valor del umbral dependerá del número de reincidencias que intentemos predecir:

$$p(j \leq 3) = 1/(1 + e^{\dagger}(-(3{,}77 - (-{,}756 \cdot x_{\downarrow}1) - (-{,}024 \cdot x_{\downarrow}2) - (1{,}424, \cdot x_{\downarrow}3) - (2{,}462 \cdot x_{\downarrow}4) - \\ - (-{,}372 \cdot x_{\downarrow}5) - (-{,}177 \cdot x_{\downarrow}6) - (,466 \cdot x_{\downarrow}7) - (,887 \cdot x_{\downarrow}8)))$$

$$p(j \leq 3) = 1/(1 + e^{\dagger}(-(3{,}77 - (-{,}756 \cdot 0) - (-{,}024 \cdot 168) - (1{,}424, \cdot 1) - (2{,}462 \cdot 1) - \\ - (-{,}372 \cdot 1) - (-{,}177 \cdot 1) - (,466 \cdot 1) - (,887 \cdot 1)))$$

$$p(j \leq 3) = ,988 \tag{6}$$

$$p(j \leq 3) = 1/(1 + e^{\dagger}(-(3{,}77 - (-{,}756 \cdot 1) - (-{,}024 \cdot 168) - (1{,}424, \cdot 1) - (2{,}462 \cdot 1) - \\ - (-{,}372 \cdot 1) - (-{,}177 \cdot 1) - (,466 \cdot 1) - (,887 \cdot 1)))$$

$$p(j \leq 3) = ,975 \tag{7}$$

Ejemplo 2:

Cálculo de la probabilidad de 3 o más reincidencias para una menor infractora de 14 años (codificada como 1), en igualdad de condiciones:

$$p(j \leq 3) = 1/(1 + e^{\dagger}(-(3{,}77 - (-{,}756 \cdot 1) - (-{,}024 \cdot 168) - (1{,}424, \cdot 1) - (2{,}462 \cdot 1) - \\ - (-{,}372 \cdot 1) - (-{,}177 \cdot 1) - (,466 \cdot 1) - (,887 \cdot 1)))$$

$$p(j \leq 3) = ,975 \tag{8}$$

Tabla 3. Estimaciones de los parámetros de la ecuación de regresión ordinal de la V.D. Número de reincidencias en función de las V.I. del modelo final

		Estimación	Error típ.	Wald	gl	Sig.	Intervalo de confianza 95%	
							Lím. inferior	Lím. superior
Umbral	[Nº_reincidencias = 0]	1,084	,982	1,220	1	,269	-,840	3,008
	[Nº_reincidencias = 1]	2,358	,979	5,804	1	,016	,440	4,277
	[Nº_reincidencias = 2]	3,340	,979	11,641	1	,001	1,421	5,259
	[Nº_reincidencias = 3]	3,777	,981	14,814	1	,000	1,854	5,700
	[Nº_reincidencias = 4]	4,162	,985	17,848	1	,000	2,231	6,093
	[Nº_reincidencias = 5]	4,591	,992	21,414	1	,000	2,647	6,536
	[Nº_reincidencias = 6]	5,137	1,007	25,999	1	,000	3,162	7,111
	[Nº_reincidencias = 7]	5,632	1,031	29,860	1	,000	3,612	7,652
	[Nº_reincidencias = 8]	5,925	1,051	31,798	1	,000	3,866	7,984
	[Nº_reincidencias = 9]	6,338	1,090	33,805	1	,000	4,201	8,474
	[Nº_reincidencias = 10]	7,037	1,200	34,373	1	,000	4,685	9,390
	[Nº_reincidencias = 11]	7,732	1,395	30,740	1	,000	4,999	10,466
Ubicación	Género	-,756	,218	12,020	1	,001	-1,183	-,328
	Edad primera infracción	-,024	,005	22,426	1	,000	-,034	-,014
	Déficit habilidades sociales	1,424	,252	31,853	1	,000	,930	1,919
	Impulsividad	2,462	,264	86,913	1	,000	1,944	2,979
	Ansiedad	-,372	,169	4,829	1	,028	-,705	-,040
	Cambio de provincia	-,177	,089	3,924	1	,048	-,351	-,002
	Consumo sustancias psicoactivas	,466	,083	31,401	1	,000	,303	,629
	Expulsiones/despidos	,887	,287	9,569	1	,002	,325	1,449

Función de vínculo: Logit.

La elevada probabilidad de reincidencia, se debe en ambos casos a que el coeficiente de la variable *Impulsividad* es muy elevado en relación al resto (2,462) y posee una alta significación estadística; indicando que el efecto de la variable *Impulsividad* sobre la reincidencia continuada es altamente significativo, y además el valor de Wald (86,913), es el más alto de la ecuación de regresión logística, lo que la señala como la variable más influyente en la reincidencia**.**

Los 2 ejemplos anteriores corresponden al modelo completo, variando únicamente el género. Ahora bien, en función de las variables que se consideren, deberemos acudir a la tabla correspondiente y sustituir los valores de cada coeficiente *j* por sus correspondientes valores de variable, teniendo en cuenta que siempre deben ser significativos. Este ejemplo planteado como supuesto nos lleva a pensar que a medida que se incrementa el número de reincidencias, las diferencias de género se desvanecen a consecuencia de un deterioro personal grave, en el que inciden las tres variables con el valor del test de Wald más alto (*Wald > 30*): *Impulsividad, Consumo de sustancias* y *Déficit en habilidades sociales*, desapareciendo casi por completo cualquier factor de protección. En cambio, es necesario hacer notar los coeficientes negativos (ver Tabla 3), *Ansiedad* (a mayor ansiedad, menor probabilidad de reincidir) y *Cambio de provincia,* que disminuyen las probabilidades de reincidir.

Como podrá deducirse, los valores probabilísticos de la reincidencia pueden modificarse, estableciendo resultados claramente diferenciados, en función del nº de reincidencias y de las variables que queramos estimar.

El modelo final obtenido incluye las variables independientes que, de acuerdo con el modelo de Andrews y Bonta (2006), pueden clasificarse en:

- Variables estáticas (por orden de poder predictivo): *Género*, *Edad de la primera Infracción, Expulsiones/despidos* y *Cambio de provincia.*
- Variables dinámicas (más susceptibles de tratamiento personalizado): *Impulsividad, Déficit en habilidades sociales, Consumo de sustancias psicoactivas y Ansiedad*.

Como conclusión, debería establecerse un sistema de seguimiento personalizado para cada niño, de tipo multidisciplinar, que apoyara y tratase a los niños con alto riesgo de reincidencia; este aspecto está contemplado en la ley del menor, pero no se ha implementado de manera eficaz.

REFERENCIAS

Agresti, A. (2002). Categorical data analysis (2nd ed.). New York: John Wiley & Sons.

Andrews, D. A. y Bonta, J. (2006). The psychology of criminal conduct (4ª. Ed.). Cincinnati: Anderson Publishing Company.

Lemert, T.E. (1967). Human deviance, Social problems and Social control. Englewood Cliffs.:Prentice Hall.

Long, J. S. & Freese, J. (2006). Regression models for categorical dependent variables using Stata (2nd ed.). Texas: Stata Press.

PROPIEDADES MÉTRICAS DEL EFECTO ESCOLAR: MAGNITUD Y CONSISTENCIA

Elsa Peña-Suárez y Ángela Campillo-Álvarez
Universidad de Oviedo
Correo electrónico: penaelsa@uniovi.es

Resumen

El objetivo del presente estudio fue medir la magnitud y la consistencia de los efectos escolares a través de las actitudes y capacidades que conforman la competencia científica. Por una parte las actitudes medidas se refieren al interés y apoyo a la investigación científica, por otra las capacidades analizadas son identificar cuestiones científicas, explicar fenómenos científicos y utilizar pruebas científicas. Se entiende por efecto escolar el porcentaje de variación en el rendimiento del alumnado debido a las características y procesos del centro en el que está escolarizado. Una vez comprobado tal efecto, se estima la consistencia de los efectos mediante la correlación entre distintas medidas criterio. En este estudio se comparan medidas cognitivas y actitudinales controlando características sociodemográficas del alumnado y del centro evaluado en PISA 2006. Los resultados indicaron un mayor efecto del centro en las medidas de capacidad, en torno al 16% de varianza explicada por el nivel dos. Así como una consistencia alta entre las capacidades medidas y baja entre las actitudes. En conclusión, el centro escolar contribuye al rendimiento en la competencia científica medida a través del programa PISA.

La línea de investigación sobre Eficacia Escolar surgió de la necesidad de demostrar que los centros educativos tenían alguna influencia sobre el desarrollo de sus alumnos. Actualmente estimar la magnitud de los efectos escolares no es suficiente, por lo que se han comenzado a estudiar las propiedades que subyacen a estos efectos tales como: Consistencia, estabilidad, eficacia diferencial y continuidad. El trabajo que se presenta se sitúa en la estimación de la magnitud y consistencia de los efectos escolares.

MAGNITUD DE LOS EFECTOS ESCOLARES

Se entiende por *efecto escolar* la capacidad de los centros educativos para influir en los resultados de sus alumnos. Es decir, es el porcentaje de varianza del rendimiento del alumno debido a las variaciones entre escuelas (Murillo, 2005).

Los resultados obtenidos varían de manera considerable de un estudio a otro. Así, Mortimore et al. (1988) estimaron que el tamaño del efecto de centro es del 10%; Mandeville y Anderson (1987) encontraron que los estudiantes adquirían prácticamente todo el conocimiento de Matemáticas en el período de escolarización a diferencia de rendimiento en Lengua o en Arte, las cuales estaban influidas por otros factores ajenos al ambiente escolar como aquéllos procedentes del entorno familiar.

En fechas recientes, Teddlie, Reynolds y Sammons (2000) realizaron un gran esfuerzo por recoger y sistematizar los resultados de 26 investigaciones más relevantes sobre la magnitud de los efectos escolares. Concluyeron que la magnitud de los efectos escolares se encuentra en torno al 15%, con una amplia variación entre países y estudios.

Con población española Murillo (2004) encuentra en centros de primaria un efecto de 9,26% en Matemáticas y de 3,37% en Ciencias Naturales entre otros. En estudios transnacionales como «*Programa para la Evaluación Internacional del Alumnado»* (PISA), el 13,9% de la varianza total está explicada por factores de contexto y proceso del centro para España. Martín et al. (2008) en un estudio longitudinal y multinivel en población española de secundaria eleva el efecto a un 20%.

Los Modelos Multinivel (MM) han suscitado gran interés en el ámbito de la IEE, dado que este tipo de análisis posibilita detectar diferentes fuentes de variación a través de los distintos niveles. Precisamente la detección de esas fuentes de variación entre los niveles de análisis es la forma de estimar la magnitud de los efectos escolares.

CONSISTENCIA DE LOS EFECTOS ESCOLARES

Podría ocurrir que en un determinado centro los alumnos que tengan buenos resultados en Matemáticas y no tan buenos en Comprensión Lectora, o que éstos tengan alta autoestima pero con nulos conocimientos de Inglés. Es lo que se llama, consistencia de los efectos escolares. Esta cuestión tiene un alto interés académico y práctico: Si los efectos escolares no son consistentes, complica el cálculo de la magnitud del efecto y, con ello, la búsqueda de escuelas eficaces y la determinación de factores asociados (Murillo, 2007).

La mayoría de las investigaciones miden la consistencia de los efectos escolares entre distintos productos por medio de la estimación de la correlación entre la eficacia de un centro obtenida a partir de dos variables de producto. Concretamente se elaboran los residuales de cada variable de producto - la diferencia entre el logro esperado y el obtenido teniendo en cuenta características contextuales del alumnado y del centro. Posteriormente se calcula el coeficiente de correlación de Pearson o el coeficiente Kappa, el cual mide el grado de discrepancia entre índices (e.g. Mandeville y Anderson, 1987; Crone et al., 1994). Existe una diversidad de resultados para estimar los efectos escolares, como se va a exponer.

ENTRE MEDIDAS COGNITIVAS

En general se puede decir que la consistencia entre áreas académicas en Primaria es moderada y positiva. Por ejemplo Sammons, Nuttall y Cuttance (1993) encontraron una correlación entre Lectura y Matemáticas de 0,61. En España Murillo (2004) observó correlaciones moderadamente altas (0,82) entre Ciencias Sociales y Naturales y en torno a 0,71 entre Matemáticas con Ciencias Naturales y Sociales. En Lengua y Matemáticas, Murillo (2007) encontró una correlación moderada de 0,50.

Para Secundaria destaca el estudio de Thomas, Sammons y Mortimore (1997) donde se analiza la consistencia de los efectos para siete áreas a lo largo del tiempo, encontrando una correlación baja y moderada, entre 0,24 y 0,72. En España Castejón (1994) obtiene similares resultados, de hecho la relación entre Matemáticas e idioma extranjero fue de 0,63 y entre Historia e idioma extranjero fue de 0,63.

Esa disparidad de resultados en los estudios comentados referidos a población española se debe no sólo a causas de que el alumnado de referencia está en distintas etapa educativa, sino también, a que los *outputs* de Murillo (2004) proceden de la aplicación de pruebas estandarizadas de rendimiento mientras que Castejón (1994) se basa en calificaciones escolares otorgadas por docentes (Murillo, 2007).

ENTRE MEDIDAS NO COGNITIVAS

La evidencia sobre la consistencia de los efectos escolares entre medidas de rendimiento socio-afectivo es más escasa. Mortimore y cols. (1988) encontraron una relación de 0,33 entre actitud hacia la Lectura y las Matemáticas; Fitz-Gibbon (1991) halló que la consistencia entre la actitud hacia las Matemáticas y la actitud hacia la escuela era de 0,48. Murillo (2007) encontró una correlación baja pero significativa (en torno a 0,32) entre convivencia social y satisfacción con el centro.

ENTRE MEDIDAS COGNITIVAS Y NO COGNITIVAS

La discrepancia mayor se encuentra cuando se trata de medir la relación entre variables cognitivas y no cognitivas. Reynolds (1976) y Rutter et al. (1979) encontraron una alta correlación entre Eficacia Escolar y Eficacia Social (asistencia y delincuencia). Sobre todo se encuentra una relación fuerte entre el ámbito cognoscitivo y afectivo cuando son áreas con cierta *«competencia afectiva»* como Música, Arte etc. (Kochan et al, 1996). Sin embargo Gray et al. (1983), Mortimore et al. (1988), Fitz-Gibbon (1991), y Murillo (2004 y 2007) concluyen que los efectos escolares medidos como resultados académicos son independientes de los estimados en variables afectivas y sociales.

En general existe una falta de estudios en población española sobre el tema que se está abordando, tanto en general como particularmente en Educación Secundaria. Asimismo no existe una evidencia concluyente de la consistencia de los efectos

en pruebas estandarizadas de rendimiento y, menos aún, en pruebas que evalúen competencias afectivas o conductas escolares. El trabajo que se presenta se sitúa en la medición de las dos propiedades descritas de los efectos escolares en la adquisición de competencias. Concretamente la competencia medida es la competencia científica, evaluada a través del programa PISA desarrollado por la OCDE, por lo que este trabajo presenta unas estimaciones de los efectos escolares fiables, que proceden de pruebas estandarizadas, validadas y adaptadas a los distintos países que participan en dicho programa.

El objetivo planteado fue estimar la magnitud y la consistencia de los efectos escolares en las actitudes y capacidades medidas en la competencia científica.

Las hipótesis planteadas:

Hipótesis uno: El tamaño del efecto se sitúa en torno al 15% de la varianza explicada.

Hipótesis dos: La consistencia de los efectos escolares es moderada en actitudes y capacidades.

MÉTODO

Participantes

La muestra estuvo compuesta por 17. 528 alumnos, cuya media de edad fue 15,84 años (0,29). La mitad de la muestra eran hombres y la otra mitad mujeres. Participaron un total de 612 centros: 56,5% son públicos; 37,28% se sitúan en ciudad grande (de 100.000 a 1.000.000 habitantes).

Variables

Las variables se situaron en una estructura jerárquica donde el nivel uno son los estudiantes y el nivel dos son los centros escolares.

Variables de alumnado: sexo; nacionalidad; repetidor; índice Socio-Económico y Cultural del alumnado (ISEC); la media de los cinco valores plausibles en cada actitud (interés en la ciencia y apoyo a la investigación científica) y en cada capacidad evaluada (identificar, explicar y utilizar conocimientos y procedimientos científicos)

Variables de centro: ISEC centro (media aritmética del ISEC familiar)

Análisis de datos

Todos los análisis fueron llevados a cabo mediante el programa HLM 7 y el SPSS 15. Para la estimación de la magnitud de los efectos escolares se calculó la

Correlación Intraclase (IC), es decir, el porcentaje de varianza explicada por el centro. Para la consistencia se calculó la diferencia entre el logro esperado y el obtenido teniendo en cuenta características contextuales, obteniéndose un residual EB (Estimadores empírico de Bayes) para cada capacidad y actitud, los cuales posteriormente se correlacionaron.

RESULTADOS

Magnitud

El IC tomó un valor mayor entre las capacidades que entre las actitudes evaluadas, tabla 1. De hecho, dicho índice tomó valores en torno al 16% de varianza explicada por el nivel de centro en las capacidades de identificar y utilizar conocimientos científicos, mientras que en las actitudes IC toma valores más residuales, en torno al 7,29%. Por lo que se evidencia una mayor magnitud de los efectos escolares en la adquisición de capacidades.

Tabla 1. Estimación del tamaño de los efectos de centro

	Actitudes		**Capacidades**		
	Interés	**Apoyo**	**Identificar**	**Explicar**	**Utilizar**
U0	569,4	676,7	1177,7	1359,5	1501,1
R	7792,6	8039,8	6209,7	7658,3	7858,9
total	8362	8716,5	7387,4	9017,8	9360
IC	0,1	0,9	0,2	0,1	0,2
Efecto de centro	6,8%	7,8%	15,94%	15,1%	16,0%

Consistencia

Las correlaciones entre los distintos residuales mostraron relaciones altas entre las capacidades evaluadas. Mientras que dichas correlaciones entre las capacidades y actitudes son bajas, como ocurre entre el interés científico y la capacidad de identificar conceptos científicos, tabla 2. Por lo que los efectos de centro se mantienen en el desarrollo de las distintas capacidades.

Tabla 2. Consistencia de los efectos de centro

	Interés	Apoyo	Identificar	Explicar	Utilizar
Interés					
Apoyo	0,59				
Identificar	0,08	0,27			
Explicar	0,11	0,29	0,82		
Utilizar	0,09	0,28	0,83	0,89	

DISCUSIÓN

Estimar la magnitud y la consistencia de los efectos escolares en las actitudes y capacidades que conforman la competencia científica en el programa PISA ha sido la finalidad de este trabajo. En general los datos obtenidos apoyan las hipótesis planteadas. Los efectos de los centros escolares encontrados en la adquisición de capacidades son los que se planteaban en la primera hipótesis y son muy similares a los obtenidos en diversos estudios (Marchesi y Martínez-Arias, 2006; Pajares-Box, 2005 y Teddlie, Reynolds y Sammons, 2000). Por lo que respecta a las variables actitudinales muestran un efecto de centro residual como se encuentran en Fitz-Gibbon (1991) y Murillo (2007).

La consistencia de los efectos escolares obtenida entre las diferentes medidas valida parcialmente la segunda hipótesis. Tal y como se planteaba los centros escolares presentan una mayor consistencia de los efectos entre las capacidades, pero se esperaba una mayor correlación entre las actitudes y entre las actitudes y capacidades evaluadas. Esta baja relación también se encuentra en otros trabajos (Knuver y Brandsma, 1993; Mortimore et al., 1988 y Murillo, 2004 y 2007, Thomas, Sammons y Mortimore, 1997). A modo de síntesis se puede afirmar que los centros educativos tienen un papel significativo en la adquisición de la competencia científica, sobre todo en el aprendizaje de conocimientos relacionados con dicha disciplina. En las actitudes no se evidencia dicho efectos de forma clara. No obstante es necesario profundizar en el estudio de los efectos escolares en pruebas actitudinales.

REFERENCIAS

Castejón, J. L. (1994). Estabilidad de diversos índices de eficacia de centros educativos. *Revista de Investigación Educativa, 24*, 45-60.

Crone, L. J., Lang, M., Franklin, B. J., y Hallbrook, A. (1994b). Composite versus component score: Consistency of school effectiveness classification. *Applied Measurement in Education 7*(4), 303-321.

Fitz-Gibbon, C. T. (1991). Multilevel modelling in an indicator system. En S. W. Raudenbush y J. D. Willms (Eds.), *Schools classrooms and pupils: international studies of schooling from multilevel perspective* (pp. 67-83). San Diego, CA: Academic Press.

Gray, J., McPherson, A. E., y Raffe, D. (1983). Reconstructions of Secundary Education: theory, myth and practice since the war. London Routledge and Kegan Paul.

Jencks, C. S., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., et al. (1972). *Inequality: a reassesment of the effect of family and schooling in America*. New York: Basic Books.

Kochan, S. E., Tashakkori, A., y Teddlie, C. (1996). You can't judge a high school by test data alone: Constructing an alternative indicator of secondary school effectiveness. Comunicación presentada en la reunión anual de American Educational Research Association. Nueva York.

Mandeville, G. K., y Anderson, L. W. (1987). The stability of school effectiveness indices across grade level and subjects areas. *Journal of Educational Measurement, 24*(3), 203-216.

Mandeville, G. K., y Anderson, L. W. (1987). The stability of school effectiveness indices across grade level and subjects areas. *Journal of Educational Measurement, 24*(3), 203-216.

Martín, E., Martínez -Arias, R., Marchesi, A., y Pérez, E. M. (2008). Variables the predict academic achievement in the spanish compulsory secondary educational system: a longitudinal multi-level analysis. *The Spanish Journal of Psychology, 11*, 400-413.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., y Ecob, R. (1988). The effects of scool membership on pupil´s educational outcomes. *Research Papers in Education, 3*(1), 3-26.

Murillo, J. (2005). *La investigación sobre eficacia escolar*. Barcelona: Octoedro.

Murillo, J. (2004). Aportaciones de la investigación sobre eficacia escolar. Un estudio multinivel sobre los efectos escolares y los factores de eficacia de los centros docentes de primaria en España. Universidad Complutense de Madrid, Madrid.

Murillo, J. (2007). *Investigación iberoamericana sobre eficacia escolar.* Bogotá: Convenio Andrés Bello.

Reynolds, D. (1976). The delinquent school. En P.Woods (Ed.). *The process of schooling* (pp. 124-178). London: Routledge Kegan Paul.

Rutter, M., Mortimore, P., Ouston, J., y Maughan, B. (1979). *Fifteen thousand hours*. Londres: Open Books.

Sammons, P., Nuttall, D. L., y Cuttance, P. (1993). Differential school effectiveness: results from reanalysis of the inner London Education Authority´s Junior School Project Data. *British Educational Research Journal, 4*, 381-405.

Teddlie, C., Reynolds, D., y Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. En C. Teddlie y D.Reynolds (Eds.), *The International Handbook of School Efectiveness Research* (pp. 55-133). London: Falmer Press.

Thomas, S., Sammons, P., y Mortimore, P. (1997). Stability and consistency in Secundary school's effects on student's GCSE outcomes over three years. *School Effectiveness and School Improvement, 8*(2), 169-197.

MÁQUINAS DE BOLTZMANN COMO ALTERNATIVA EN LA IMPUTACIÓN DE DATOS

Sergio M. Vergara, Manuel J. Sueiro e Iván Sánchez-Iglesias
Universidad Complutense de Madrid
Correo electrónico: msueiro@psi.ucm.es

Resumen

Las omisiones de respuesta distorsionan los resultados de cuestionarios y tests. Como método para solventar este problema, este trabajo propone utilizar una Máquina de Boltzmann como procedimiento para imputar valores perdidos y compara el funcionamiento de este método con otras técnicas clásicas de imputación de datos, como las basadas en la media, *two-way*, una adaptación del cálculo de las probabilidades esperadas como alternativa a la imputación de datos, *response function* y las redes de Hopfield. Se realizó un estudio de simulación donde el número de sujetos (500, 1000 y 2000), la longitud del test (10, 20 y 40 ítems) y el porcentaje de valores perdidos (1%, 5% o 10%) se manipularon para comprobar el rendimiento de los diferentes procedimientos de imputación. Así mismo se utilizaran datos perdidos simulados completamente al azar (MCAR) y perdidos al azar (MAR), para comprobar el funcionamiento de dicha red ante distintos tipos de valores perdidos. Los resultados obtenidos muestran que los datos imputados mediante las Máquinas de Boltzmann mantienen el grado de acierto de los sujetos y el valor de la respuesta, mientras que técnicas como *two-way* mantienen las distribuciones de los datos originales.

Frente al problema de los valores perdidos, Rubin (1976) desarrolló la tipología aún vigente para diferenciar entre los distintos tipos de valores perdidos según las distintas relaciones probabilísticas entre los valores completos (Y_{com}), valores perdidos (Y_{mis}) y los valores observados (Y_{obs}): perdidos al azar (*missing at random:* MAR), cuando la probabilidad de ser perdidos dependen de los valores observados, pero no de los valores perdidos; perdidos totalmente al azar (*missing completly at random:* MCAR), cuando no sólo no dependen de los valores observados sino que su ocurrencia tampoco depende de los valores perdidos; y perdidos no al azar (*missing not at random:* MNAR), cuando su ocurrencia no depende de los valores observados y sin embargo sí dependen de los valores perdidos.

Los métodos de imputación de datos tienen como objetivo el sustituir los valores perdidos de cualquiera de los distintos tipos (MAR, MCAR y MNAR) por valores que se aproximen lo más posible a lo que el sujeto hubiese respondido. Sin

embargo, los métodos de imputación son más plausibles para valores MAR y MCAR, ya que al ser valores con un alto contenido de azar y, al no tener reglas desconocidas en los patrones, se pueden aproximar los valores de manera estocástica (Schafer & Graham, 2002).

Métodos basados en la media. Basado en métodos de imputación clásicos como es la imputación por la media (IM) (Schafer & Graham, 2002), Bernaards and Sijtsma (2000) desarrollaron un método de imputación llamado *two-way* (TW), mediante el uso combinado de otros métodos de imputación siguiendo la expresión:

$$TW_{ij} = PM_i + IM_j - OM; TW_{ij} \in \Re$$

donde PM_i es el Pearson Mean (Huisman, 1998) y OM es el Overall Mean, es decir, la media en el ítem de todos los sujetos que lo han respondido.

El método *probabilidades esperadas* (PE) es un método propuesto por nosotros, similar al TW pero de manera multiplicativa. Partiendo de las tablas de contingencia, se calcula la probabilidad esperada de acertar un ítem. Las puntuaciones marginales para el cálculo son proporcionales al IM y el PM mencionados anteriormente.

Es trivial demostrar que la expresión para la obtención de esta probabilidad esperada queda reducida a:

$$PE_{ij} = \frac{IM_i PM_j}{OM}$$

Métodos basados en modelos no paramétricos. *Response Function Imputation* (RF) (Sijtsma & Van der Ark, 2003) es un método basado en el contexto de la Teoría del Respuesta al Ítem para modelos no-paramétricos y en el cálculo del valor Restscore (Junker & Sijtsma, 2000).

Métodos basados en redes neuronales. Las redes neuronales han demostrado su utilidad en distintos campos. En concreto, dentro del problema de la imputación de datos se ha propuesto el uso de las Redes Neuronales de Hopfield (Hopfield & Tank, 1982) y, para solucionar sus limitaciones, las máquinas de Boltzmann (Acley & Hinton,1985).

La arquitectura de las Redes Neuronales de Hopfield (RH) se compone de una única capa de neuronas interconectadas de pesos continuos, simétricos y no autorrecurrentes. Se basa en modelos de excitación-inhibición. La red trata de estabilizar dichas conexiones mediante una función decreciente de energía que sigue la siguiente expresión:

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j + \sum_i \theta_i s_i$$

siendo w_j el peso de la conexión entre las neuronas i y j; s_i la componente i del vector de entrada y θ_i el umbral asociado a la neurona i.

La limitación que presentan este tipo de redes, es que su comportamiento es determinista, tendiendo a replicar los patrones que tiene almacenados y, en el caso de ser un patrón nuevo, generando una salida lo más parecida al patrón más cercano.

Como evolución de las RH, proponemos el uso de las máquinas de Boltzmann, que comparten arquitectura y función de energía con las RH pero cuya función de activación se basa en la siguiente regla estocástica (Acley & Hinton, 1985):

$$p_{j'} = \frac{1}{(1 + e^{-\Delta E_j / T})}$$

siendo $p_{j'}$ la probabilidad de que se active la unidad *j'* y *T* la temperatura de la red.

La temperatura es la que modifica el comportamiento estocástico de la red: si el valor de *T* es elevado la red tiene un comportamiento aleatorio, mientras que si es bajo, la red se comporta de manera determinista, como una red de Hopfield. Hinton, G. & Sejnowski, T (1986) recomiendan utilizar templado simulado, un procedimiento consistente en comenzar con valores altos de T y disminuirlos progresivamente, para evitar el problema de los mínimos locales.

La imputación mediante redes neuronales parte del concepto de que cada unidad de la red representa un ítem y cada patrón de activación de la red es el patrón de respuestas de un sujeto. La red se entrena con los sujetos que no tienen valores perdidos para obtener los pesos de las conexiones. Después se presentan los patrones incompletos de manera que la red debe reconstruirlos, fijando la activación de las unidades que representan ítems con valores conocidos y dejando variar a activación de las unidades que representan valores perdidos.

Este trabajo propone utilizar una Máquina de Boltzmann como procedimiento para imputar valores perdidos y compara el funcionamiento de este método con otras técnicas de imputación de datos: IM, TW, PE, RF y RM.

MÉTODO

Factores bajo estudio. Se valoraron dos tipos de valores perdidos (Schafer y Graham, 2002): valores perdidos completamente al azar (MCAR) y valores perdidos al azar (MAR). Bajo la condición MCAR, la probabilidad de un valor estuviera perdido era igual para todos los sujetos. En el caso MAR, la probabilidad de tener un valor perdido de los sujetos con valor en una covariable $Y = 1$ era mayor (el doble) que en el grupo $Y = 2$.

Las probabilidades de valor perdido utilizadas fueron 0,01; 0,05 y 0,1, para muestras de 500, 1000 y 2000 sujetos y tests de 10, 20 y 40 ítems.

Generación de datos. Se generaron 100 réplicas para cada condición experimental. Para cada sujeto, se generó una covariable binaria *Y* (Van Ginkle & Van der Ark, 2007). Los niveles del rasgo de los sujetos, θ, se generaron siguiendo la dis-

tribución normal con desviación típica 1 y media $\mu_1 = -0.25$ para $Y = 1$ y $\mu_2 = 0.25$ para $Y = 2$. Las respuestas dicotómicas de los sujetos se simularon bajo el modelo logístico de dos parámetros.

Indicadores valorados. Se utilizaron dos tipos de indicadores para medir el grado de eficacia de los distintos métodos. Los indicadores locales realizan la valoración del comportamiento de cada ítem, y se obtuvieron mediante la proporción de proporción de imputaciones correctas a nivel de ítem y a nivel de la puntuación total en el test. Los indicadores globales valoran la integridad de los datos en la muestra. El sesgo en la varianza y sesgo en alfa de Cronbach se utilizaron para esto.

Métodos de Imputación. Los métodos de imputación utilizados fueron IM, TW, PE, RF, RH y MB. Fueron programados en Free-basic y analizados con el programa SPSS 19.

RESULTADOS

En la Tabla 1 se muestra la proporción de imputaciones correctas. Los métodos basados en las redes neuronales son los que mejores resultados presentan en las diferentes condiciones. MB obtuvo las proporciones de acierto más altas, con unos resultados bastante estables a través de todas las condiciones. RF fue el método no basado en redes neuronales con mejores tasas de acierto, aunque su eficacia decrece con respecto al resto de métodos cuando hay 40 ítems y un 10% de valores perdidos.

En términos generales hay poco efecto de las condiciones experimentales, manteniéndose estables los resultados incluso en las condiciones más restrictivas.

La Tabla 2 nos muestra los resultados del sesgo en el alfa. En este caso, en general los distintos métodos producen algo de sesgo en el alfa, siendo RF el que mejor se comporta en las condiciones con menor número de valores perdidos (0.05 y 0.1). TW y PE tienen comportamientos en general con poco sesgo, aunque parecen funcionar mejor en tests largos.

Los métodos basados en las redes neuronales, MB y RH tienden a infraestimar el resultado del índice de fiabilidad (alfa). En general, el sesgo es mayor en las imputaciones realizadas mediante las máquinas de Boltzmann.

Tabla 1. Proporción de imputaciones correctas

Nº Ítems	N	P (Perdido)	MCAR						MAR					
			MB	RH	IM	TW	PE	RF	MB	RH	IM	TW	PE	RF
10	500	,01	,75	,71	,60	,67	,68	,70	,76	,70	,61	,68	,68	,68
		,05	,75	,70	,60	,68	,67	,69	,75	,70	,59	,68	,67	,70
		,10	,75	,70	,61	,67	,67	,69	,75	,70	,60	,68	,68	,70
	2000	,01	,76	,70	,60	,69	,68	,70	,76	,70	,60	,68	,68	,70
		,05	,75	,70	,61	,68	,67	,70	,76	,70	,60	,68	,68	,70
		,10	,75	,70	,61	,68	,67	,69	,75	,70	,60	,68	,68	,69
40	500	,01	,78	,72	,59	,69	,68	,71	,78	,73	,59	,69	,68	,72
		,05	,78	,72	,61	,69	,68	,65	,78	,72	,60	,69	,69	,66
		,10	,72	,69	,60	,69	,68	,50	,77	,71	,60	,69	,68	,53
	2000	,01	,79	,72	,60	,69	,68	,72	,79	,72	,60	,69	,68	,72
		,05	,79	,73	,61	,69	,69	,72	,79	,73	,60	,69	,69	,72
		,10	,77	,71	,60	,69	,68	,57	,78	,71	,60	,69	,69	,67

Tabla 2. Sesgo en Alfa (en milésimas)

Nº Ítems	N	P (Perdido)	MCAR						MAR					
			MB	RH	IM	TW	PE	RF	MB	RH	IM	TW	PE	RF
10	500	,01	-2	-1	4	-1	-1	0	-2	-1	5	0	-1	0
		,05	-12	-6	24	-3	-3	0	-13	-7	24	-4	-4	0
		,10	-22	-8	53	-7	-6	0	-22	-10	51	-7	-7	0
	2000	,01	-2	-1	5	-1	0	0	-3	-2	4	-1	-1	0
		,05	-12	-6	25	-4	-3	0	-12	-5	25	-3	-3	0
		,10	-23	-13	53	-8	-7	0	-23	-13	51	-7	-7	0
40	500	,01	-1	-1	2	0	0	0	-1	-1	2	0	0	0
		,05	-3	-2	9	1	1	4	-3	-2	9	0	1	4
		,10	0	4	20	1	2	23	-5	-1	19	1	1	22
	2000	,01	-1	-1	2	0	0	0	-1	-1	2	0	0	0
		,05	-3	-2	9	0	1	0	-3	-2	9	0	1	0
		,10	-5	-2	20	1	2	16	-7	-4	19	1	1	8

DISCUSIÓN

El presente trabajo trata de mostrar la utilidad del uso de las redes neuronales en el problema de la imputación de datos, como alternativa a las técnicas más clásicas (IM) y modernas (TW, RF y RH). También proponemos el uso de otra técnica desarrollada por nosotros, PE, que se fundamenta en cálculos computacionalmente más sencillos y que, a su vez, hace uso de otros métodos de imputación. Los resultados han mostrado que ambos métodos propuestos son buenas alternativas, en función de cuales sean los objetivos finales con el conjunto de datos.

En general, la longitud del test y la probabilidad de valor perdido fueron factores clave (mostrando un peor desempeño conforme aumentaban estas variables),

mientras que el tamaño de la muestra apenas tuvo efecto. Respecto al tipo de valores perdidos, el método RF funcionó mejor en MAR que en MCAR, mientras que el resto obtuvieron un desempeño similar bajo ambas condiciones.

Los resultados obtenidos mediante los indicadores locales nos muestran que los métodos basados en las redes neuronales, en particular MB, ofrecen mejores resultados que el resto de métodos. El componente estocástico de MB, junto al procedimiento de templado simulado, evita a la red caer en mínimos locales, lo que permite recuperar mejor el patrón de respuestas original que utilizando RH.

Los resultados obtenidos mediante los indicadores globales, muestran un mejor funcionamiento por parte de los métodos basados en combinar la media del ítem con la media de la persona: TW y la nueva propuesta PE. En este ámbito, el funcionamiento de la máquina de Boltzmann fue aceptable, aunque inferior al de los citados. Esto indica que existe un sesgo en las imputaciones incorrectas al usar redes neuronales.

En función de para qué queramos utilizar los datos, unos métodos de imputación se muestran más eficaces que otros. Así, si lo que queremos es mantener la distribución de los datos y el comportamiento conjunto de los mismos para obtener índices que tengan en cuenta el conjunto global, será recomendable usar métodos de imputación como TW o PE. Sin embargo, si lo que queremos es ver las respuestas que daría un sujeto en concreto, o el grado de acierto de ciertos ítems, se recomendaría el uso de las redes neuronales, en particular las máquinas de Boltzmann.

En estudios futuros, sería interesante usar métodos de contraste más robustos que nos permitan conocer mejor las diferencias entre las MB y el resto de métodos de imputación. La diferencia entre los parámetros reales y los estimados (Finch, 2008) podría ser una buena opción. Así mismo, se hace necesaria la extrapolación de estas técnicas a ítems politómicos.

REFERENCIAS

Ackley, D. H. y Hinton, G. E., Sejnowsk, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Science, 9*, 147-169.

Bernaards, C. A. & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*, 321-364.

Hinton, G. E. y Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. En: Rumelhart, J. L. McClelland, and the PDP group (Eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, (282-317). Cambridege, MA: MIT Press.

Hopfield, J. J. & Tank, D. W. (1985). *'Neural' Computation of decisions in optimization problems.* Biological Cybernetics.

Huisman, M. (1998). Item nonresponse: Occurrence, causes and imputation of missing answers to test items. Leiden, The Netherlands*: DSWO Press.*

Junker, B. W., & Sijtsma, K (2000). Latent and manifest monotonicity in item response models. *Applied psichologycal measurement, 24,* 65-81.

Rubin, D. B. (1976). Inference and missing data*. Biometrika, 63*, 581-592.

Schafer, J. L. & Graham, J. W. (2002). Missing Data: Our View of the Stte of the Art. *Psychological Methods*, *7* (2), 147-177.

Sijtsma, K. & Van der Ark, L. A. *(2003*). Investigation and treatment of missing ítem scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505-528.

Van Ginkel, J. R., Van der Ark, A. y Sijtsma, K. (2007). Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results. *Multivariate Behavioral Research, 42* (2), 387-414.

EXPERIENCIAS DOCENTES

UN ANÁLISIS SEMIÓTICO DE RECURSOS INTERACTIVOS PARA LA ENSEÑANZA DE LA PROBABILIDAD CONDICIONAL

José M. Contreras[1], Carmen Díaz[2], Gustavo R. Cañadas[1], y Pedro Arteaga[1]
[1] Universidad de Granada
[2] Universidad de Huelva
Correo electrónico: jmcontreras@ugr.es

Resumen

La probabilidad condicional es un concepto requerido en la inferencia estadística, clásica y bayesiana, asociación entre variables, regresión, modelos lineales y toma de decisiones. Sin embargo, en la investigación didáctica se han descrito numerosos sesgos de razonamiento (Díaz y de la Fuente, 2005), entre ellos los siguientes: (a) Independencia y mutua exclusividad: Creer que dos sucesos son independientes si y sólo si son excluyentes (Kelly y Zwiers,1986); (b) Confusión entre condicionamiento y causación (Falk, 1986; Gras y Totohasina, 1995); (c) Intercambio de sucesos en la probabilidad condicional (Falk, 1986); (d) Confusión de probabilidad condicional y conjunta (Tversky y Kahneman, 1982). Respecto al razonamiento bayesiano, las limitaciones son debidas, entre otras razones, a la expresión verbal, cantidad de datos y condiciones involucradas.

En la actualidad existen muchos recursos para la enseñanza de la estadística en Internet, que permiten explorar y simular procesos aleatorios y que podrían contribuir a la mejora de la intuición del alumno y por tanto a la superación de algunos de los sesgos descritos. Sería necesario, sin embargo, realizar análisis didácticos de estos recursos, así como de la forma de trabajo con ellos para orientar al profesorado. En este trabajo llevamos a cabo un estudio sobre algunos recursos que permiten explorar y visualizar la probabilidad simple y condicional y temas relacionados con ella, con el fin de prever su utilidad en la enseñanza y las posibles dificultades que se pueden encontrar a la hora de utilizar estos contenidos en el aula con estudiantes.

La probabilidad simple y condicional se requieren en la probabilidad producto, inferencia estadística, clásica y bayesiana, asociación, regresión, modelos lineales y toma de decisiones. Sin embargo, la investigación didáctica ha descrito sesgos de razonamiento, que continúan después de la enseñanza (Diaz y de la Fuente, 2005). Los más importantes sesgos son los siguientes:

– *Independencia y mutua exclusividad:* Creer que dos sucesos son independientes si y sólo si son excluyentes. Kelly y Zwiers (1986) suponen es debi-

do a la imprecisión del lenguaje ordinario, en que «independiente» puede significar, a veces, separado.

– *Confusión entre condicionamiento y causación:* Mientras que una relación causal implica la dependencia estadística, lo contrario no es siempre cierto. Sin embargo, la persona que evalúa una probabilidad condicional percibe en forma diferente las relaciones causales y diagnósticas, dando mayor peso a la causal (Tversky y Kahneman, 1982a). La relación de causalidad también se asocia, con la secuencia temporal suponiendo que el suceso condicionnante ha de suceder antes que el condicionado (Falk, 1986; Gras y Totohasina, 1995).

– *Intercambio de sucesos en la probabilidad condicional* (Eddy, 1982; Falk, 1986). Muchos estudiantes no diferencian *P(A/B)* y *P(B/A)*.

– *Confusión de probabilidad condicional y conjunta* (Pollatsek, Well, Konold y Hardiman, 1987; Ojeda, 1995; Tversky y Kahneman, 1982b).

– *Situaciones sincrónicas y diacrónicas:* Si el problema se plantea como una serie de experimentos secuenciales (situaciones diacrónicas) o simultáneos (sincrónicas) (Falk, 1989; Ojeda, 1995).

– *Razonamiento bayesiano*: Tversky y Kahneman, (1982b) indican que la comprensión del teorema de Bayes exige gran esfuerzo cognitivo. Gras y Totohasina (1995) suponen que los alumnos pueden encontrarse con dificultades en función del tipo de representación elegida para resolver el problema.

En la actualidad existen muchos recursos en Internet, que permiten explorar y simular procesos aleatorios y podrían contribuir a mejorar la intuición del alumno y a superar algunos de estos sesgos. En este trabajo llevamos a cabo un estudio sobre algunos de estos recursos con el fin de prever su utilidad en la enseñanza y las posibles dificultades que se pueden encontrar los estudiantes al usarlos.

EXPLORACIÓN DE OPERACIONES ENTRE SUCESOS

En primer lugar analizamos un recurso (Figura 1) que permite explorar las operaciones entre dos sucesos. Se muestra un diagrama de rectángulo con una partición del espacio muestral en un suceso *A,* su contrario y otro suceso *B* y su contrario. Las probabilidades de *A* y *B* y de sus contrarios están fijadas. Pinchando con el ratón, se puede colorear diferentes sucesos, $A, B, \overline{A}, \overline{B}, A \cap B, \overline{A \cap B}, A \cup B$ y $\overline{A \cup B}$. La posición relativa de los sucesos *A* y *B* se pueden modificar, visualizando las probabilidades $P(A \cap B)$, $P(\overline{A \cap B})$, $P(A \cup B)$ y $P(\overline{A \cup B})$. Este programa calcula automáticamente las probabilidades condicionales $P(B / A)$ y $P(A / B)$.

El Applet nos muestra directamente la descomposición de la probabilidad de la intersección y la unión, probabilidad del complementario, leyes de Morgan, que permiten calcular la probabilidad del contrario de la unión e intersección y cálculo de la probabilidad condicional a partir de la intersección.

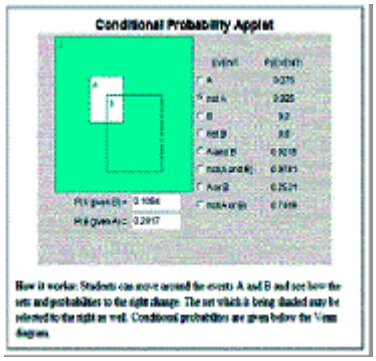


Figura 1. Pantalla del Conditional Probability Applet

Un primer objetivo es que el alumno perciba el significado de la intersección, la unión y los complementarios y de cómo cambian las probabilidades según la posición relativa de los sucesos. También permite observar la diferencia entre $P(A/B)$ y $P(B/A)$, ya que muchos estudiantes confunden estas dos probabilidades, error que Falk (1986) denomina falacia de la condicional transpuesta. Otra posible aplicación sería comprobar que independencia no es lo mismo que mutua exclusividad. Moviendo los sucesos A y B hasta que no tengan intersección común (es decir sean mutuamente excluyentes) se observa claramente que tanto la probabilidad condicional $P(A/B)$ como la de la intersección $P(A \cap B)$ son nulas, aunque el producto de las probabilidades de los sucesos A y B no lo sea. También se puede la falacia de la conjunción (Tversky y Kahneman, 1982b) o creencia de que es más probable la intersección de dos sucesos que la de uno de ellos por separado o la de su unión.

DIFICULTADES POSIBLES DE LOS ESTUDIANTES

Una de las principales dificultades que pueden encontrar los estudiantes es la interpretación del lenguaje del Applet. En la columna de la izquierda aparecen diferentes operaciones con sucesos bajo la palabra «event» y las notaciones de las operaciones con sucesos, aunque intuitivas son correctas. Sin embargo, en la columna derecha sólo aparece mención a P(event), pero luego en cada fila no vuelve a aparecer la notación de probabilidad. Es por ello que los estudiantes podrían considerar todas las probabilidades listadas como probabilidades simples (en lugar de referirse a la probabilidad de la unión, intersección o contrario). La notación coloquial para los sucesos intersección y la probabilidad condicional puede también ocasionar el error pues Einhorn y Hogarth (1986) sugieren que los enunciados que usan la conjunción «y» pueden llevar a confundir la probabilidad conjunta y la probabilidad condicional. Por otro lado, como no se puede cambiar el tamaño relativo de los sucesos A y B, se puede interpretar que la probabilidad solo depende de la posición relativa de los sucesos, aunque en realidad también dependería del tamaño de los sucesos en relación al espacio muestral.

EXPLORACIÓN DE OTROS CONCEPTOS

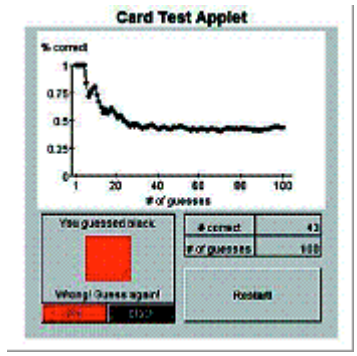


Figura 2. Card test Applet

En la figura 2 mostramos un recurso que puede servir para explorar la idea de independencia. Los estudiantes deben imaginar una baraja de cartas con tarjetas rojas y negras y prededir la ocurrencia de diferentes sucesos. El porcentaje de tarjetas rojas puede ser modificado. Se trata de concienciar a los estudiantes de que la probabilidad de cada suceso no varía en función de los resultados obtenidos. Se les debe alentar a jugar el juego de varias maneras. En primer lugar, hacemos un ejercicio para estimar la proporción de tarjetas rojas encubiertas o para adivinar el porcentaje de cartas de color negro. Los estudiantes deben ser animados a reflexionar sobre la idea de independencia y sobre la existencia de sesgos tales como la falacia del jugador. También se puede hacer observar la estabilización de las frecuencias relativas a la larga, pero hacer notar las fluctuaciones en las series cortas de ensayos.

El recurso mostrado en la figura 3 consta de un cuadrado de seis por seis que representa las 36 posibilidades a la hora de tirar dos dados de seis caras y nos permite investigar cómo se comportan las probabilidades condicionales. Podemos hacer una elección entre dos listas que aparecen en la pantalla para visualizar los diferentes sucesos que se presentan en el Applet y decidir cual es el suceso condicionante. Cuando se haya elegido el suceso de cada lista, algunas de las celdas del cuadrado se colorearan de rojo o amarillo. Los cuadrados de color representan el número de combinaciones de los dados que satisfagan la condición B (la segunda condición de la derecha de la lista desplegable). De estas celdas de color, el rojo representa las combinaciones, que también cumplen la primera condición (A). Existen dos métodos para calcular $P(A/B)$, uno de ellos implica contar los cuadrados de colores, el otro utiliza una fórmula. El Applet nos proporciona métodos para ver cómo se relacionan entre sí. Si pulsamos «Reverse», se intercambian las declaraciones A y B. Con lo que debemos detectar rápidamente que $P(A/B)$ no es, en general, igual a $P(B/A)$. Finalmente, en la tabla 1. presentamos las direcciones de estos y otros recursos de exploración.

Tabla 2. Recursos para visualización de la probabilidad condicional y conceptos relacionados

Nombre	Dirección
Bayes Rule	www.bolderstats.com/gallery/prob/bayes.html
Birthday Demonstration	onlinestatbook.com/simulations/birthday/birthday.html
Card test Applet	http://www.stat.tamu.edu/~west/Applets/cardtest.html
Condicional probability	onlinestatbook.com/chapter5/conditional_demo.html
Conditional probability	www.rfbarrow.btinternet.co.uk/htmasa2/Prob2.htm
Conditional probability and independent events	www.cut-the-knot.org/Curriculum/Probability/ConditionalProbability.shtml
Conditional Probability and Multiplication Rule	www.spsu.edu/math/deng/m2260/stat/cond/cond.html
Conditional probability	www.stat.tamu.edu/~west/Applets/Venn1.html
Conditional Probability	onlinestatbook.com/simulations/conditional_p/conditional_p.html
Dice and conditional probability	www.math.fau.edu/Richman/Liberal/dice.htm
Gamblers Fallacy	onlinestatbook.com/simulations/gambler_fallacy/gambler.html
Java Applets: TwoArm	www.dim.uchile.cl/~mkiwi/ma34a/libro/chapter4/TwoArm/TwoArm.html
Marbles	www.shodor.org/interactivate/activities/marbles/
Probabilty by Surprise	www-stat.stanford.edu/~susan/surprise/
Racing Game with One Die	www.shodor.org/interactivate/activities/RacingGameWithOneDie/
Random Birthday Applet	www-stat.stanford.edu/~susan/surprise/Birthday.html
Two Events: Conditioning	www.stat.wvu.edu/SRS/Modules/ProbLaw/GivenProb.html
Venn Conditional	www.bolderstats.com/gallery/prob/conditional.html
Venn Diagram	www.bolderstats.com/gallery/prob/venn.html
Venn Diagram Applet	www.teachers.ash.org.au/miKemath/VennDiagramApplet/VennGame.html
Venn Diagram Shape	www.shodor.org/interactivate/activities/ShapeSorter/
Venn Diagrams	www.shodor.org/interactivate/activities/VennDiagrams/
Venn Diagrams and Probability	www.stat.berkeley.edu/~stark/Java/Html/Venn3.htm

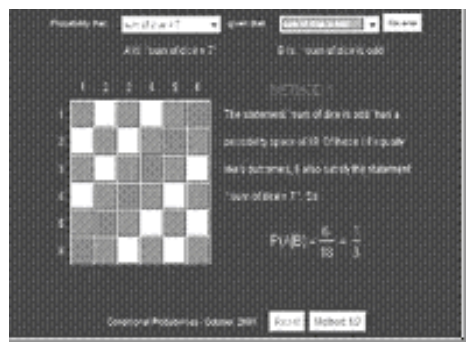


Figura 3. Conditional probability Applet

REFLEXIONES FINALES

Hemos incluido en este análisis recursos que pueden servir para visualizar algunos de los objetos matemáticos que se relacionan tanto con la probabilidad simple, como con la probabilidad condicional, o algunas de las propiedades o teoremas relacionados con los mismos. En el trabajo en el aula, se plantearía el problema, dejando un tiempo para que los estudiantes lleguen a una posible solución. Seguidamente se discutirían con los estudiantes las soluciones correctas e incorrectas encontradas por los mismos, hasta lograr que se acepte alguna de las correctas. El profesor ayudaría a analizar las causas de los errores y haría un resumen de lo aprendido. En caso de resistencia a la solución, se dejaría confrontar las soluciones con la evidencia empírica producida por el Applet para que los estudiantes comprendan las causas de sus intuiciones erróneas y las revisen. Pensamos que en este juego se dan las condiciones de idoneidad didáctica, que Godino, Wilhelmi y Bencomo (2005) definen como la articulación de seis componentes:

- Idoneidad epistémica o matemática: Representatividad de los significados institucionales implementados (o pretendidos), respecto de un significado de referencia. El proceso descrito podría se idóneo para el estudio de los conceptos de: probabilidad condicional, experimento compuesto, dependencia e independencia y experimentos dependientes e independientes, pero esta idoneidad depende del tipo de solución encontrada. En general las soluciones formales tienen mayor idoneidad en un curso universitario y de formación de profesores, pero en un curso de secundaria las soluciones intuitivas podrían ser suficientes. La solución empírica, tiene, en general, baja idoneidad matemática, a menos que se complemente con una solución intuitiva o formal.
- Idoneidad cognitiva: Grado en que los significados pretendidos/ implementados son asequibles a los alumnos, así como si los significados personales logrados por los alumnos son los pretendidos por el profesor. La situación planteada tiene suficiente idoneidad en cursos de formación de profesores de secundaria y los últimos cursos de secundaria, pues los razonamientos descritos están al alcance de los alumnos.
- Idoneidad interaccional: Grado en que la organización de la enseñanza permite identificar conflictos semióticos y resolverlos durante el proceso de instrucción. Este tipo de idoneidad dependerá de cómo organiza el profesor el trabajo en el aula. Será importante que los estudiantes trabajen en grupos para que surja el conflicto y se explicite. Será importante también organizar una discusión colectiva de las soluciones para que los mismos alumnos ayuden a sus compañeros a detectar los puntos equivocados.
- Idoneidad mediacional: Disponibilidad y adecuación de los recursos necesarios para el desarrollo del proceso de enseñanza-aprendizaje. No se precisa de muchos recursos, pues incluso podría hacerse una simulación con objetos físicos o con un solo ordenador en el aula, donde los alumnos pueden jugar colectivamente.

– Idoneidad emocional: Interés y motivación del alumnado en el proceso de estudio. Pensamos que esta es la más alta de todas, pues el juego interesa a todo el que trata de resolverlo.

En los cursos de formación de profesores, el análisis didáctico, similar al descrito, sirve para aumentar el conocimiento de los profesores sobre probabilidad, metodología de la enseñanza de la probabilidad y algunos razonamientos erróneos de los estudiantes. Se podría mejorar el proceso si se dispone de soluciones dadas por alumnos reales que los profesores puedan analizar para detectar los errores descritos.

NOTA DE LOS AUTORES

Agradecimiento a Proyecto EDU2010-1494; Beca FPU-AP2009-2807 (MCINN-FEDER); becas FPU-AP2007-03222 y FPI BES-2008-003573 (MEC-FEDER) y grupo FQM126 (Junta de Andalucía).

REFERENCIAS

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. En D. Kahneman, P. Slovic y Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Einhorn, H. J. y Hogart, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99*, 3 – 19.

Díaz, C. y de la Fuente, I. (2005). Razonamiento sobre probabilidad condicional e implicaciones para la enseñanza de la estadística. *Epsilon, 59*, 245-260.

Falk, R. (1986). Conditional Probabilities: insights and difficulties. En R. Davidson y J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics*. (pp. 292 – 297). Victoria, Canada: International Statistical Institute.

Godino, J., Wilhelmi, M. y Bencomo, D. (2005). Suitability criteria of a mathematical instruction process. A teaching experience of the function notion. *Mediterranean Journal for Research in Mathematics Education,* 4(2), 1-26.

Gras, R. y Totohasina, A. (1995) Chronologie et causalité, conceptions sources d'obstacles épistémologiques à la notion de probabilité conditionnelle. *Recherches en Didactique des Mathématiques, 15*(1), 49-95.

Kelly, I. W. y Zwiers, F. W. (1986). Mutually exclusive and independence: Unravelling basic misconceptions in probability theory. *Teaching Statistics, 8*, 96 – 100.

Ojeda, A. M. (1995) Dificultades del alumnado respecto a la probabilidad condicional. *UNO, Nº 5*, 37-55.

Pollatsek, A., Well, A. D., Konold, C. y Hardiman, P. (1987). Understanding conditional probabilities. *Organization, Behavior and Human Decision Processes, 40*, 255 – 269.

Tversky, A. y Kahneman, D. (1982a). Causal schemas in judgment under uncertainty. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 117 – 128). Cambridge, MA: Cambridge University Press.

Tversky, A. y Kahneman, D. (1982b). On the psychology of prediction. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 69 – 83). Cambridge, MA: Cambridge University Press.

VALORACIÓN DEL GRADO EN DIRECCIÓN Y GESTIÓN PÚBLICA DESDE LA PERSPECTIVA DEL ALUMNADO

Pilar García-Soidán y Xosé Mahou-Lago
Universidad de Vigo
Correo electrónico: xmahou@uvigo.es

Resumen

La transformación de la Diplomatura en Gestión y Administración Pública en el Grado en Dirección y Gestión Pública, en la Facultad de Ciencias Sociales y de la Comunicación de la Universidad de Vigo, supuso diseñar un nuevo título adaptado a las directrices establecidas en el proceso de convergencia al Espacio Europeo de Educación Superior (EEES). A petición del alumnado, se planteó la implantación en bloque de los 4 cursos que integraban el mencionado Grado en el curso académico 2009/2010, dando la opción de continuación a aquellos estudiantes que decidiesen proseguir con el plan de estudios de la Diplomatura. Todo esto supuso un esfuerzo de coordinación por parte del profesorado y un esfuerzo de adaptación del alumnado para su incorporación a una titulación que iniciaba su andadura. Por ello, en este trabajo se ha tratado de conocer el grado de valoración del proceso de implantación del Grado, desde la perspectiva del alumnado.

En el curso 2009/2010 comenzó a impartirse el Grado en Dirección y Gestión Pública en la Facultad de Ciencias Sociales y de la Comunicación de la Universidad de Vigo. La planificación de esta titulación tomó como referente la Diplomatura en Gestión y Administración Pública, que se venía impartiendo en la Universidad de Vigo, según el plan de estudios aprobado en el año 1999. La nueva titulación tenía como objetivo proporcionar una formación, tanto de carácter teórico como técnico en aspectos concretos de gestión administrativa y financiera, a los futuros cuadros intermedios de las Administraciones públicas (Grupo B). El *Consello de Goberno* de la Universidad de Vigo aprueba el 5 marzo de 2008 las «Directrices propias sobre estructura y organización académica de los planes de estudios de grado» en las que adapta la normativa estatal sobre EEES al ámbito concreto de esta universidad, además de fijar el procedimiento y calendario de elaboración de los diferentes planes de estudio.

Siguiendo este procedimiento, la *Xunta de Centro* de la Facultad de Ciencias Sociales y de la Comunicación delega, a principios de 2008, a la *Xunta de Titulación* la reforma de la Diplomatura de Gestión y Administración Pública, constituyéndose en Comisión el 31 de marzo. El resultado de un año de deliberaciones

entre las diferentes áreas da lugar a un plan de estudios de marcado carácter multidisciplinar y altamente novedoso al introducir, por primera vez, dentro del sistema universitario español un título reglado que contempla competencias propias de la Dirección Pública. Este hecho no solo motiva el cambio de denominación, sino que lleva a reorganizar el conjunto de asignaturas ya existentes y a diseñar nuevos contenidos como respuesta a la demanda de formación de los nuevos perfiles profesionales que fueron surgiendo en los últimos años en las organizaciones públicas estatales, autonómicas, locales e incluso comunitarias.

El Grado en Dirección y Gestión Pública abarca 4 cursos académicos repartidos en 8 semestres. En cada semestre, el alumnado debe superar 30 créditos ECTS, o 5 asignaturas de 6 créditos ECTS cada una (excepto en el 2º semestre del 2º curso en donde se contemplan 2 asignaturas de 9 créditos). Además de créditos vinculados a asignaturas, los/as estudiantes deben realizar, en 4º curso, un Trabajo Fin de Grado obligatorio (12 ECTS) y pueden llevar a cabo prácticas externas (12 ECTS). En total, el número de créditos a cursar es de 240, distribuidos, según el tipo de asignatura, de la forma indicada en la tabla 1.

Tabla 1. Distribución de créditos ECTS por tipo de asignatura

Tipo de asignatura	Nº créditos ECTS a cursar	Nº créditos ECTS ofertados
Formación básica	60	60
Formación obligatoria	120	120
Formación optativa	36	84
Trabajo Fin de Grado	12	12
Prácticas externas	12	12
Total	**240**	**288**

Las materias que configuran el plan de estudios se distribuyen en cinco bloques: político-administrativo, económico, jurídico, social e instrumental.

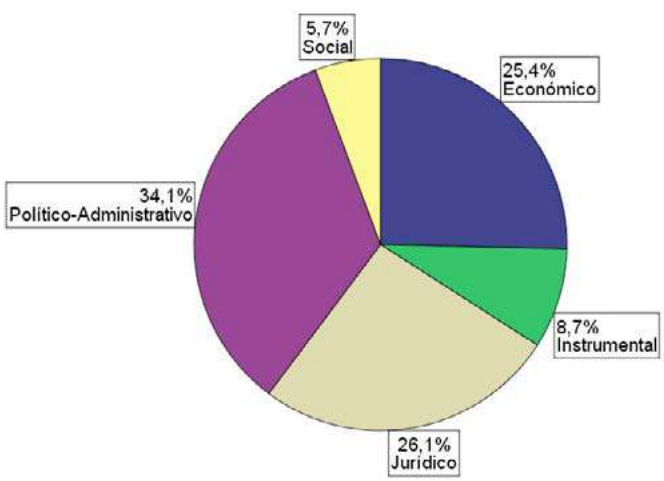


Figura 1. Distribución de créditos ECTS por bloques de materias

La implantación del Grado se realizó en bloque a petición de la mayoría del alumnado matriculado en la Diplomatura en el año académico anterior. Sin embargo, debía darse la opción de continuación a aquellos estudiantes que decidiesen proseguir con el plan de estudios de la Diplomatura. De este modo, comenzaron a impartirse simultáneamente las materias de los 4 cursos que configuraban el Grado y, a su vez, se integraron en ellas las asignaturas de los 2 últimos cursos de la Diplomatura. Todo ello supuso un esfuerzo de coordinación por parte del profesorado para la puesta en marcha del plan de estudios y, paralelamente, un esfuerzo de adaptación del alumnado para su incorporación a una titulación que iniciaba su andadura y que, en el caso de los estudiantes que no eran de nuevo ingreso, se llevó a cabo mediante un sistema de validaciones. Teniendo presente esta situación, a finales del curso académico 2009/2010 se realizó una encuesta entre el alumnado matriculado en el Grado, para conocer su grado de valoración del proceso de implantación de dicha titulación, cuyos resultados se resumen en el presente trabajo.

MÉTODO

Diseño del estudio

Para desarrollar este estudio, se diseñó una encuesta cuyas características se resumen en la ficha técnica.

Participaron en dicho estudio 91 estudiantes distribuidos por curso como se presenta a continuación (se indica el curso más alto en el que estaban matriculados).

Tabla 2. Ficha técnica del estudio cuantitativo

Ámbito:	Grado en Dirección y Gestión Pública.
Universo:	Alumnado de los 4 cursos de la titulación en el curso 2009/2010.
Tamaño de la muestra:	91 alumnos/as.
Procedimiento de muestreo:	Muestreo aleatorio simple.
Error de muestreo:	Para un nivel de confianza del 95% y P=Q, el grado máximo de error es del 5,8% para el conjunto de esta titulación.
Método de entrevista:	Encuestas realizadas mediante entrevista personal.
Período de realización:	Mayo-Septiembre de 2010.

Tabla 3. Distribución por curso

Curso	Nº alumnos/as	% alumnos/as
1º	25	27,5%
2º	21	23,1%
3º	9	9,9%
4º	36	39,5%
Total	**91**	**100%**

PERFIL DEL ALUMNADO PARTICIPANTE EN EL ESTUDIO

El gráfico siguiente refleja la distribución por género de los estudiantes entrevistados.

La mayoría de los estudiantes habían iniciado sus estudios en la Diplomatura en Gestión y Administración Pública, según se refleja en el gráfico siguiente.

Desde la implantación de estos estudios, se han distinguido dos perfiles de alumnado, según realizaran o no alguna actividad profesional, destacando como grupo predominante entre los primeros los trabajadores de la Administración Pública. Para la muestra analizada, se observa que algo más de la tercera parte de los encuestados compatibilizan trabajo y estudios**.**

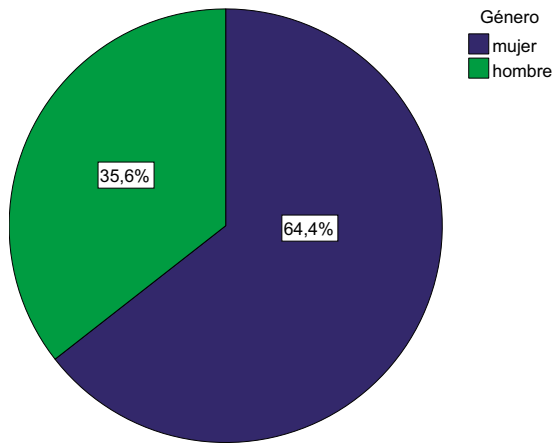


Figura 2. Distribución por género

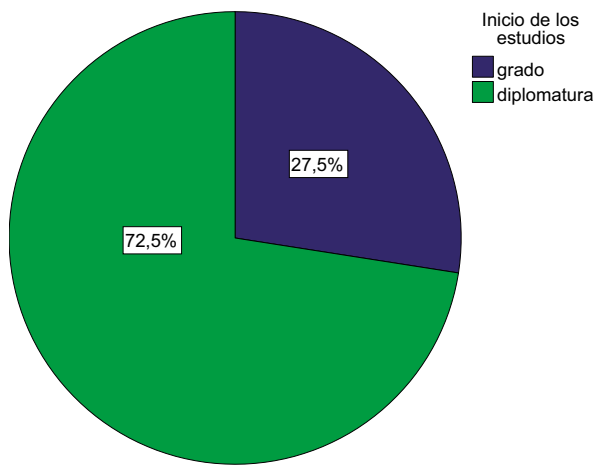


Figura 3. Distribución según la titulación inicial

Cabe destacar que con la implantación del Grado se ha reducido el porcentaje de estudiantes que desarrollan alguna actividad profesional remunerada, pasando del 40,9% entre el alumnado que proceden de la Diplomatura al 28% entre los que iniciaron sus estudios directamente en el Grado. Aunque la disminución señalada no resulta significativa (test chi-cuadrado, p-valor=0,256), cabe tener presente que la nueva titulación supone cursar un año académico adicional, que incrementa las dificultades de compatibilización de trabajo y estudios.

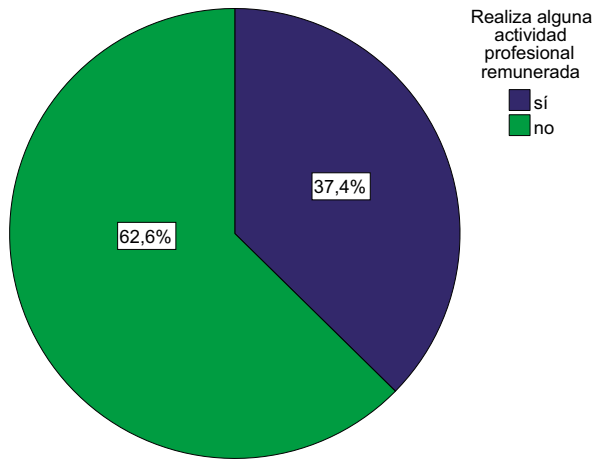


Figura 4. Distribución según actividad profesional

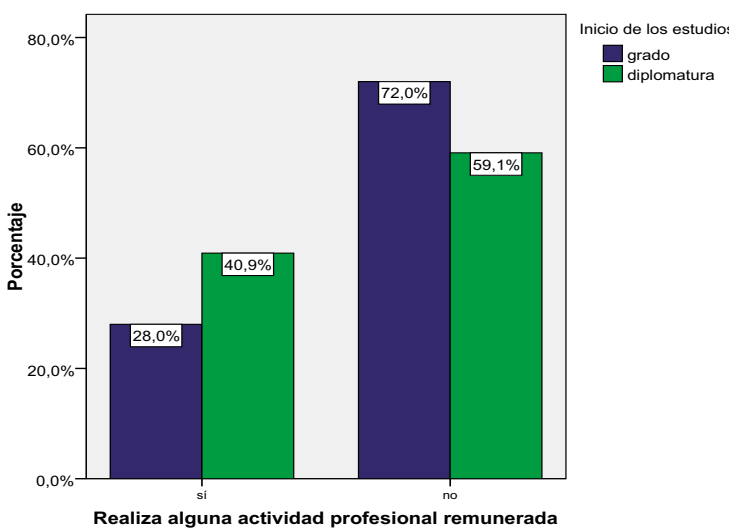


Figura 5. Distribución según actividad profesional y titulación inicial

PRINCIPALES RESULTADOS Y DISCUSIÓN

Mediante una serie de preguntas, se trató de conocer la valoración del alumnado sobre distintos aspectos del proceso de enseñanza-aprendizaje. En particular, se sometió a consideración de los participantes la idoneidad de los horarios que debían cursar. De la información obtenida, reflejada a continuación para el conjunto de la muestra, se observa que no es negativa en general, aunque el grado de descontento alcanza un porcentaje importante (41,1%).

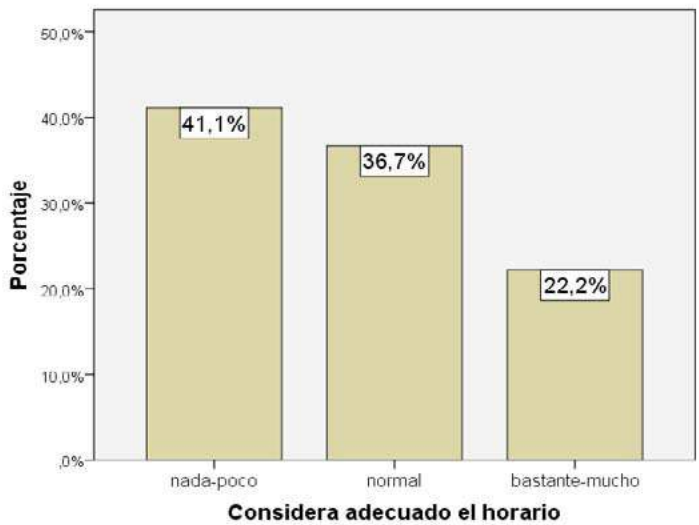


Figura 6. Valoración del horario

Las dos titulaciones (Grado y Diplomatura) se han impartido siempre en horario de tarde, para permitir la incorporación de las personas que trabajan en turno de mañana (situación predominante en el sector público). A este respecto, se observan diferencias significativas sobre la valoración del horario según la realización o no de actividad profesional (test chi-cuadrado, p-valor=0,013), resultando claramente más crítica la opinión de los estudiantes con dedicación exclusiva, de modo que más del 50% consideran poco o nada adecuados los horarios.

También influye en la opinión sobre los horarios establecidos la titulación inicial de estudios (test chi-cuadrado, p-valor=0, 017), obteniéndose una valoración negativa para el 60% de las personas que comenzaron directamente en el Grado, frente al 33,85% de los que iniciaron sus estudios en la Diplomatura.

Otro aspecto sobre el que se indagó fue la percepción general que el alumnado tenía sobre la metodología docente, que se resume en la figura 9.

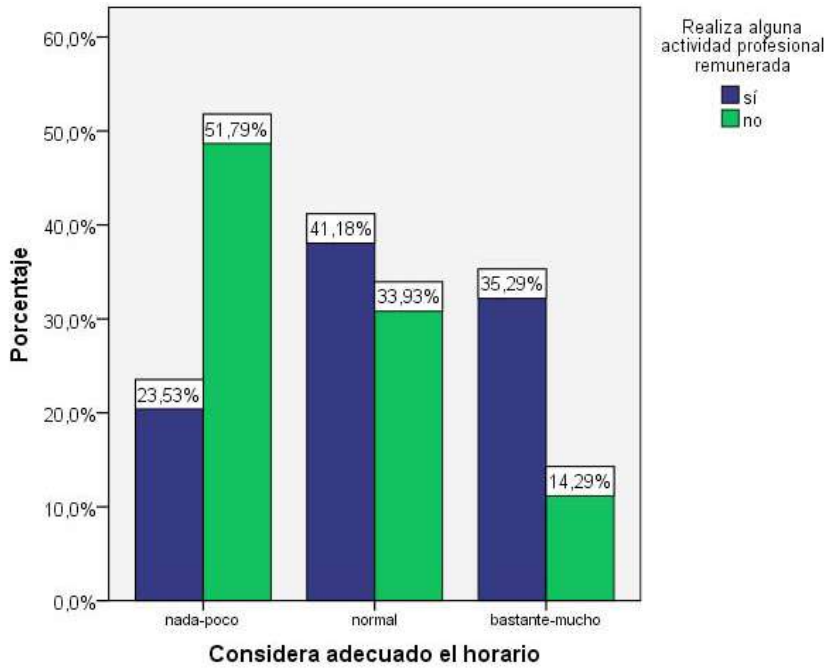


Figura 7. Valoración del horario según actividad profesional

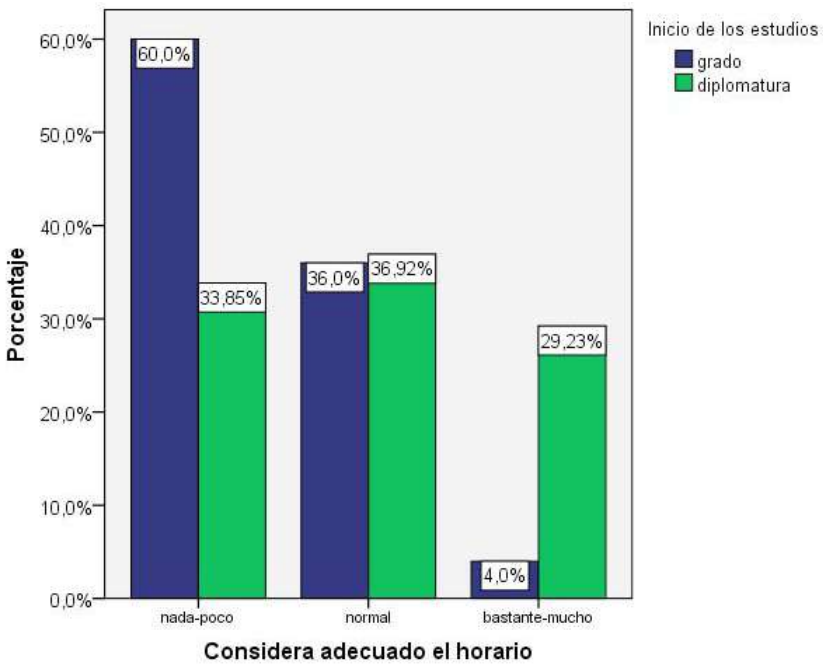


Figura 8. Valoración del horario según titulación inicial

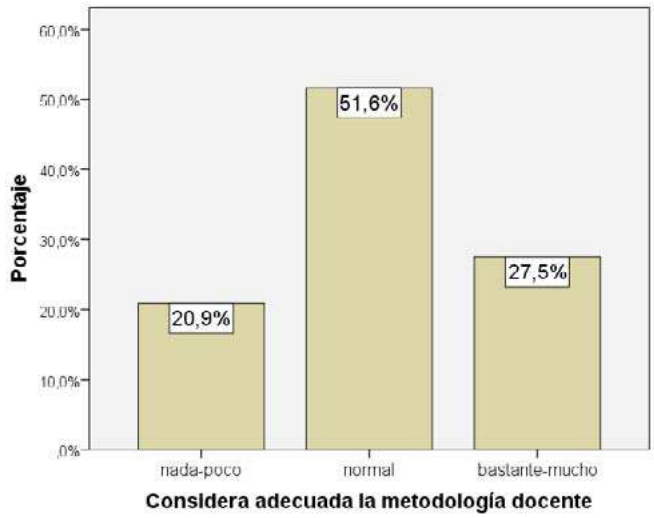


Figura 9. Valoración de la metodología docente

Se aprecia que las valoraciones no son negativas, en general, para el conjunto de la muestra. Sin embargo, al disgregar por inicio de estudios, se muestran significativamente más críticas las personas que comenzaron directamente en la titulación de Grado (test chi-cuadrado, p-valor=0,018), alcanzando el nivel de descontento el 40% entre este alumnado, frente al 13,64% entre los estudiantes provenientes de la Diplomatura.

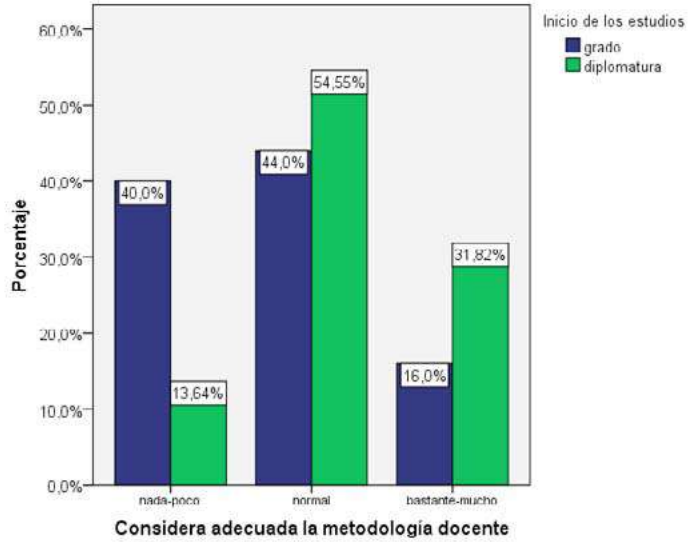


Figura 10. Valoración de la metodología docente por titulación inicial

Se preguntó también a los estudiantes sobre el sistema de evaluación, de modo que casi las tres cuartas partes lo califican como normal o positivo.

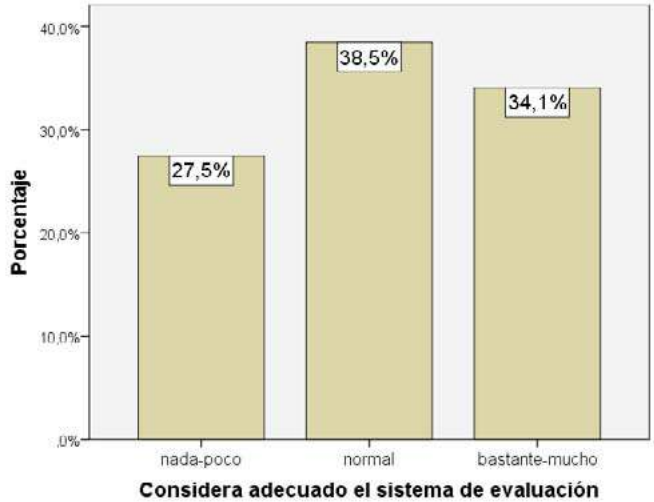


Figura 11. Valoración del sistema de evaluación

De nuevo, al analizar la información obtenida según la realización de actividad profesional (test chi-cuadrado, p-valor=0,035), resulta significativamente más desfavorable la valoración del sistema de evaluación para los estudiantes con dedicación exclusiva, entre los cuales la tercera parte lo califica como nada o poco adecuado.

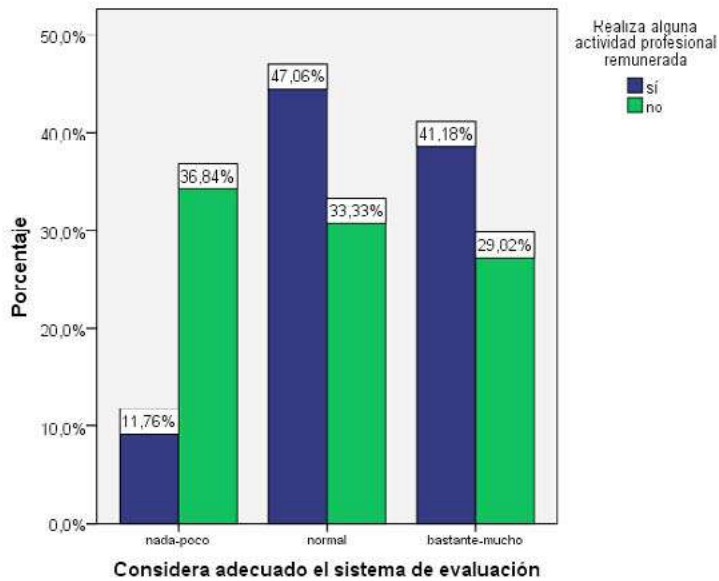


Figura 12. Valoración del sistema de evaluación según actividad profesional

Finalmente, pareció interesante conocer en qué medida, desde la perspectiva del alumnado, la formación recibida estaba orientada al empleo, que ha sido una de las premisas del proceso de convergencia de las titulaciones al EEES. Salvo para un porcentaje inferior al 20% de encuestados, parece que la docencia avanza en la dirección adecuada para responder a las demandas laborales. No se observaron diferencias significativas para esta valoración con respecto a la titulación inicial o la realización de actividad profesional.

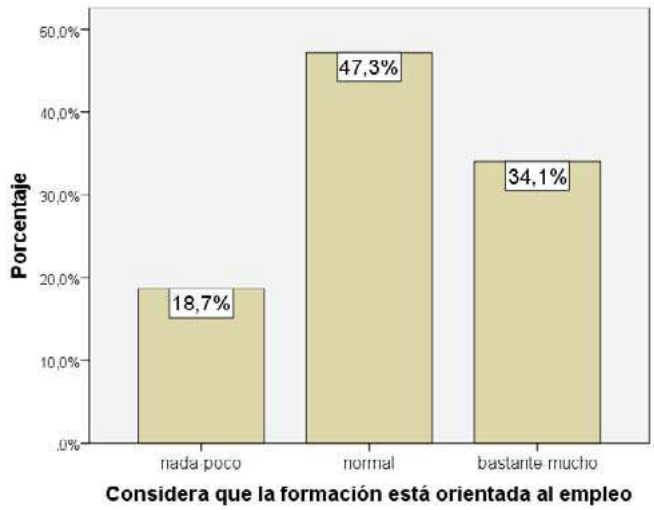


Figura 13. Valoración de la orientación de la formación para el empleo

CONCLUSIONES

La implantación del Grado supuso introducir diversos cambios en el proceso de enseñanza-aprendizaje, con respecto a la titulación anterior, cuyo efecto se ha tratado de medir en el presente trabajo desde la perspectiva del estudiante. Entre otras innovaciones, se han incorporado tres modalidades de grupos docentes en función del número de alumnos/as (grande, intermedio y reducido) y se ha dado un peso específico al trabajo autónomo del estudiante (docencia no presencial). Todo ello ha repercutido en la organización de la docencia presencial a través de los horarios, la metodología y el sistema de evaluación. Sobre estos aspectos, la opinión del conjunto de la muestra es, en general, satisfactoria, aunque conviene matizar algunas diferencias.

Para la mayoría de los estudiantes la distribución horaria resulta adecuada, si bien está condicionada por dos factores, la carrera inicial elegida (Diplomatura o Grado) y el hecho de compatibilizar trabajo y estudios. El alumnado que comenzó sus estudios directamente en la titulación de Grado es más crítico en la valoración de los horarios, que aquellos estudiantes que iniciaron sus estudios en la Diploma-

tura. Esta circunstancia resulta especialmente sorprendente si se tiene en cuenta que los alumnos que cambiaron de carrera podían estar matriculados simultáneamente en materias de cursos distintos, por efecto del sistema de validaciones y, en consecuencia, deberían afrontar, a priori, una distribución horaria más complicada. Una posible explicación a esta diferencia de valoración según la carrera inicial elegida, estaría en que el alumnado procedente de la Diplomatura demandó activamente, incluso a través de acciones de protesta, la implantación en bloque del Grado, creando una ambiente favorable hacia la nueva titulación.

Por otro lado, la realización de una actividad profesional implica una valoración de los horarios más satisfactoria que la que manifiestan las personas que no trabajan, pudiendo estar relacionado este hecho con la franja horaria de tarde reservada a la titulación de Grado. Además, las personas que desarrollan una actividad laboral (en su mayoría como empleados de la Administración Pública) están motivadas por el hecho de aprender y disponer de recursos que les faciliten el ascenso en la escala profesional.

La percepción general del alumnado sobre la metodología docente o el sistema de evaluación toma valores en torno a una adecuación normal-buena. Sin embargo, se aprecian diferencias significativas en la valoración de la metodología, según la titulación inicial elegida, y en la consideración sobre el sistema de evaluación, en función de que se realice o no alguna actividad laboral. En este sentido, aportan una visión más crítica acerca de la metodología docente las personas que comenzaron directamente sus estudios en la titulación de Grado. Por otra parte, el alumnado con dedicación exclusiva a sus estudios considera menos adecuado el mecanismo de evaluación que las personas que desempeñan algún trabajo.

En lo que se refiere a la preparación para la salida al mundo laboral, la opinión del alumnado es satisfactoria, considerando que la formación recibida está orientada al empleo y no se observa ningún factor que marque diferencias significativas en estas valoraciones.

REFERENCIAS

Aldecoa, F. (Coord.) (2005). *Libro Blanco del Título de Grado de Ciencia Política y de la Administración, Sociología y Gestión y Administración Pública*. Madrid: Aneca.

Cao Abad, R., Francisco Fernández, M., Naya Fernández, S., Presedo Quindimil, M.A., Vázquez Brage, M., Vilar Fernández, J.A. & Vilar Fernández, J.M. (2001). *Introducción a la Estadística y sus aplicaciones*. Pirámide.

Ruíz Díaz, M.A. & Pardo Merino, A. (2002). SPSS 11. *Guía para el análisis de datos*. McGraw-Hill/interamericana de España.

Santos Peñas, J., Muñoz Alamillos, Á., Juez Martel, P. & Cortiñas Vázquez, P. (2003). *Diseño de encuestas para estudios de mercado. Técnicas de Muestreo y Análisis Multivariante.* Madrid: Centro de Estudios Ramón Areces.

Pérez Fernández, J.M. (Coord.) (2003). *IV Encuentro CIGAP: Conferencia Interuniversitaria de diplomaturas en Gestión y Administración Pública*. Oviedo: Universidad de Oviedo.

Directrices propias da Universidade de Vigo sobre estrutura e organización académica dos plans de estudo de Grao (aprobado en Consello de Goberno, 5 de Marzo de 2008).

RD 1426/1990, de 26 de Octubre, por el que se establece el título universitario oficial de Diplomado en Gestión y Administración Pública y las directrices generales propias de los planes de estudios conducentes a la obtención de aquél (BOE número 278 de 20/11/1990, páginas 34358-34359).

RD1393/2007, de 29 de Octubre, por el que se establece la ordenación de las enseñanzas universitarias oficiales (BOE número 260 de 30/10/2007, páginas 44037-44048).

Resolución de 15 de octubre de 2010, de la Universidad de Vigo, por la que se publica el plan de estudios de Graduado en Dirección y Gestión Pública (BOE número 264 de 1/11/2010, páginas 91893-91896).

ADAPTANDO «DISEÑOS DE INVESTIGACIÓN» AL GRADO DE PSICOLOGÍA: SEGUIMIENTO DE LAS INNOVACIONES DOCENTES DESDE LA PERSPECTIVA DE LOS ESTUDIANTES

Olatz López Fernández, Manel Viader Junyent, Antoni Cosculluela Mas, Mª Luisa Honrubia Serrano, Joan Mª Malapeira Gas, Lucia Pirla Buil, Nuria Aparicio Lopez y Lorena Manzano Diaz
Universidad de Barcelona
Correo electrónico: olatzlopez@ub.edu

Resumen

Se describe la adaptación de la asignatura de «Diseños experimentales y aplicados» de la Facultad de Psicología de la Universidad de Barcelona en el marco de las directrices del Espacio Europeo de Educación Superior. Los cambios realizados han sido la estructuración del contenido en tres modalidades según el plan docente europeo (presenciales, semipresencial y no presencial), la implementación del campus virtual Moodle con un diseño basado en la actividad del estudiante, así como una metodología de evaluación continuada basada en un sistema de pruebas de validación. Los resultados han sido obtenidos a través de una adaptación del cuestionario EMID «*Evaluación del Modelo de Innovación Docente*», que permite evaluar un modelo de innovación docente en función de una serie de dimensiones. El objetivo de este trabajo es evaluar la gestión que los estudiantes realizan de su trabajo académico en DEIA utilizando el campus virtual como complemento a las clases presenciales, así como conocer su percepción hacia la innovación de la docencia dirigida a la adaptación de la asignatura a la convergencia europea. Finalmente, entre las conclusiones halladas destaca la valoración positiva del modelo manifestado por los estudiantes.

La asignatura troncal «Diseños experimentales y aplicados» de la Facultad de Psicología de la Universidad de Barcelona se ha ido adaptando progresivamente para encajar en el marco de las directrices del Espacio Europeo de Educación Superior (EEES) según la Universidad de Barcelona (UB) (UB, 2006a). Las adaptaciones se pueden resumir en tres: diseño de un nuevo plan docente europeo con una estructura del contenido que respeta las tres modalidades de enseñanza (sesiones presenciales, sesiones semi-presenciales con un trabajo tutorizado y actividad no presencial mediante un trabajo autónomo del estudiante) en base a la propuesta de distribución de los créditos europeos («European Credit Transfer System», ECTS) indicada en las directrices de la UB para la elaboración de los nuevos planes docentes (2006b); implementación de un entorno virtual de aprendizaje (EVA), me-

diante el campus virtual institucional, que corresponde a la plataforma de código abierto MOODLE («Modular Oriented-Object Dynamic Learning Enviorment») que adopta esta Universidad en el curso académico 2006-2007 con un estudio piloto en el que se participa (Viader, López, Rifà, Cifré, Cosculluela y Malapeira, 2007); metodología de evaluación continuada, que se inicia en el curso 2007-2008 en la UB (2006a). En DEIA está basada, el 80%, en un sistema de pruebas de validación individuales teórico-prácticas (material teórico y problemas de prácticas presenciales y autónomas) y, el 20% restante, en el trabajo tutorizado o dirigido (grupal e individual).

FASES DE IMPLANTACIÓN DEL MODELO EUROPEO (2007-2009)

Este estudio se realiza durante los cursos académicos 2007-2009 en diversos grupos de DEIA, que pertenecen al mismo equipo docente. Se imparte con los mismos profesores, las mismas fuentes bibliográficas, idéntica metodología de enseñanza, aprendizaje y evaluación, así como campus virtual (con el mismo diseño, estructura y funcionalidad) (véase figura 1).

Figura 1. Pantalla principal del campus virtual DEIA

Se diseñó un sistema de evaluación continua en base a pruebas de validación del conocimiento teórico-práctico de la asignatura. Para ello se desarrolló un nuevo sistema de actividades prácticas, donde todo el planteamiento se basaba en tres temáticas de investigación de nuestra disciplina abordables desde todos los diseños de investigación estudiados en la asignatura. Estas temáticas eran el hilo conductor de dos tipos de problemas a resolver:

Se diseñó un sistema de evaluación continua en base a pruebas de validación del conocimiento teórico-práctico de la asignatura (80% de la nota final de DEIA). Para ello se desarrolló un nuevo sistema de actividades prácticas, presenciales

(«problemas clase» y no presenciales («problemas campus»), donde todo el planteamiento se basaba en tres temáticas de investigación de nuestra disciplina abordables desde todos los diseños de investigación estudiados en la asignatura. Estas temáticas eran el hilo conductor de dos tipos de problemas a resolver:

– *Prácticas presenciales de clase*: eran grupales y no evaluables directamente.

– *Prácticas no presenciales autónomas*: eran individuales y evaluables directamente.

A parte se continuaba con el Trabajo dirigido, no presencial y tutorizado que se adaptó al sistema: se realizaba inicialmente en pequeño grupo (cuatro miembros) y la última parte de forma individual. Se evaluaba aparte (20% de la nota final de DEIA). Se estructuraba en tres actividades: búsqueda de resúmenes a través de las bases de datos científicas respecto a un tema asignado (en grupo), análisis metodológico del tema de investigación (en grupo) y análisis metodológico de un artículo científico original en inglés (individual).

Los criterios de evaluación del sistema de evaluación continua eran comunicados desde la presentación de la asignatura. Consistían en tres pruebas de validación de conocimientos teórico-prácticos, en que se sumaba la puntuación obtenida de los problemas campus si se superaba un filtro, que consistía en que el estudiante debía superar cada prueba de validación.

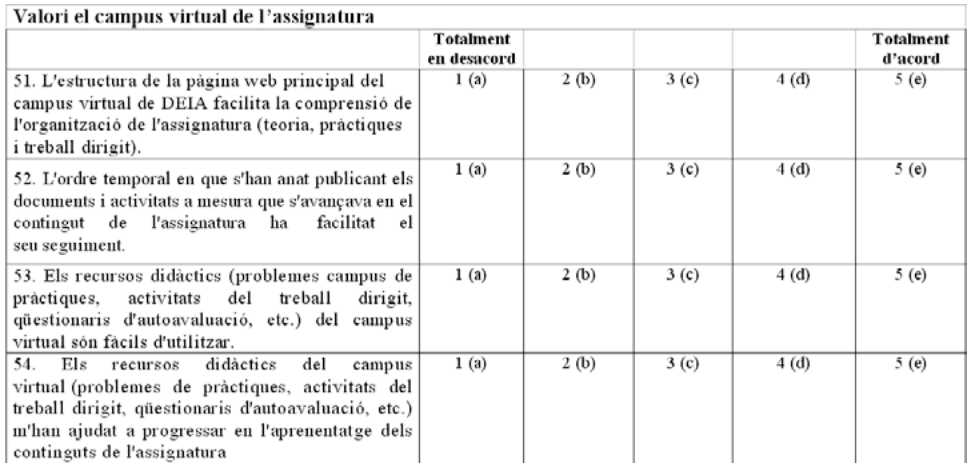
La innovación docente brevemente descrita fue valorada mediante la administración de un cuestionario para evaluar bajo el punto de vista de los estudiantes cómo percibían estos cambio, en que DEIA había sido una de las asignaturas pioneras, tanto en la adaptación del plan docente europeo como en el uso de Moodle como elemento organizador e interactivo que iba a ser la base de la actividad y evaluación de estudiante.

MÉTODO

La muestra era de 268 estudiantes de segundo curso de Psicología de la UB, de 7 grupos de DEIA, durante los cursos 2007-2008 y 2008-2009 (previos a la implantación definitiva del Grado de Psicología). Al no existir diferencias significativas en ambos cursos se analizaron conjuntamente. La mayoría eran mujeres (82.8%) no repetidoras (83.2%). El 93.6% asistía a más del 60% de las clases. El 81.2% empezaba a estudiar unos días antes de la prueba. No trabaja el 43.3%. Tenían ordenador (98.5%) e Internet en casa (94.7%).

Para el estudio de esta innovación se administró una adaptación del cuestionario EMID (Bono, Arnau y Blanca, 2005), que se re-diseñó con el fin de estudiar la innovación docente realizada en el nuevo sistema de evaluación de DEIA y el uso del campus virtual para el logro de los objetivos de aprendizaje (López-Fernández et al., en prensa).

El cuestionario se administró a los estudiantes en formato papel durante 30 minutos de la última práctica del curso (Mayo 2008 y Mayo 2009). Era anónimo y voluntario (véase figura 2).



Valori el campus virtual de l'assignatura					
	Totalment en desacord				Totalment d'acord
51. L'estructura de la pàgina web principal del campus virtual de DEIA facilita la comprensió de l'organització de l'assignatura (teoria, pràctiques i treball dirigit).	1 (a)	2 (b)	3 (c)	4 (d)	5 (e)
52. L'ordre temporal en que s'han anat publicant els documents i activitats a mesura que s'avançava en el contingut de l'assignatura ha facilitat el seu seguiment.	1 (a)	2 (b)	3 (c)	4 (d)	5 (e)
53. Els recursos didàctics (problemes campus de pràctiques, activitats del treball dirigit, qüestionaris d'autoavaluació, etc.) del campus virtual són fàcils d'utilitzar.	1 (a)	2 (b)	3 (c)	4 (d)	5 (e)
54. Els recursos didàctics del campus virtual (problemes de pràctiques, activitats del treball dirigit, qüestionaris d'autoavaluació, etc.) m'han ajudat a progressar en l'aprenentatge dels continguts de l'assignatura	1 (a)	2 (b)	3 (c)	4 (d)	5 (e)

Figura 2. Ejemplo de cuestionario EMID

RESULTADOS

Primero, se muestran los resultados psicométricos de la adaptación del EMID, que consta de 100 ítems (9 de variables sociodemográficas y 91 tipo Likert de 1 a 5) con el propósito de estudiar el impacto de la innovación en el aprendizaje de los estudiantes. Consta de 10 dimensiones: datos sociodemográficos, experiencia previa en este tipo de evaluación, experiencia previa con la utilización de las Tecnologías de la Información y la Comunicación (TICs), estimación del tiempo invertido en las actividades, valoración del sistema de evaluación continua, valoración del campus virtual, valoración del sistema enseñanza-aprendizaje-evaluación del campus virtual, valoración del sistema de enseñanza semipresencial. Se comprobó la fiabilidad y validez de esta adaptación.

Factores	Validez interna (α de Cronbach)	Porcentaje de variancia explicada
Campus virtual como gestor de aprendizaje	.953	12.174
Semipresencialidad	.928	8.695
Teoría y prácticas (actividades individuales)	.908	6.470
Campus virtual como soporte al aprendizaje	.873	4.793
Sistema de evaluación continuada	.841	4.074
Frecuencia de estudio	.756	3.311
Horas de estudio	.753	3.142
Recursos didácticos	.648	2.870
TIC	.833	2.757
Trabajo dirigido (actividades grupales)	.654	2.383

Figura 3. Propiedades psicométricas EMID adaptado.

En segundo lugar, se muestran los resultados obtenidos a partir de los apartados principales del cuestionario se revisan a continuación.

Estimación del tiempo invertido en las actividades

Los estudiantes dedican más horas al trabajo dirigido, concretamente a la consulta de Psycinfo, en comparación con el resto (estudiar, «problemas campus», complementarias, etc.). Las actividades de menor dedicación son la consulta bibliográfica y del campus virtual. En cuanto los materiales, los contenidos del campus virtual (recursos o actividades) y, sobretodo, los apuntes de clase son los más utilizados, siendo de nuevo la bibliografía la menos usada.

Valoración del sistema de evaluación continuada

Valoran muy adecuadamente el sistema de evaluación continua, especialmente las tres pruebas de validación.

Experiencia en el tipo de evaluación

Los problemas campus y los problemas clase son los que mayor puntuación obtienen. En cambio, la interacción a través del campus virtual es valorada como poco útil. Las prácticas de ordenador indiferentes.

Valoración del campus virtual

Los elementos didácticos publicados (recursos o actividades) y la evaluación continuada han ayudado en el aprendizaje de los contenidos de la asignatura. Las herramientas de comunicación de Moodle (mensajería, foros, etc.) son indiferentes.

Valoración del sistema de enseñanza-aprendizaje-evaluación

Todo es valorado positivamente siendo el campus virtual y sus contenidos considerados útiles para el aprendizaje, pero éste no motiva a los estudiantes.

Experiencia previa con el uso de las TIC

No estaban familiarizados con el aprendizaje y evaluación mediante TIC.

Valoración del sistema de enseñanza semipresencial

A pesar de que valoran positivamente esta modalidad, la consideran un complemento a las clases presenciales, que afirman que siguen siendo lo más importante.

ELEMENTOS DE LA INNOVACIÓN MÁS RELEVANTES EN FUNCIÓN DEL NIVEL DE COMPRENSIÓN DE DEIA Y DEL NIVEL DE SATISFACCIÓN CON EL SISTEMA PROPUESTO

	Comprensión			Satisfacción		
	Baja Rango medio	Alta Rango medio	*U*	Baja Rango medio	Alta Rango medio	*U*
Tiempo invertido:						
Horas bibliografía	98.61	92.52	3622.500	98.69	99.08*	2519.000
Valoración EC:						
Trabajo presencial	72.36	103.82***	2523.000	62.13	106.97***	1643.000
Trabajo autónomo	67.68	107.38***	2237.500	55.51	109.73***	1385.000
Trabajo dirigido	88.00	96.12	3450.000	73.49	104.13*	2086.000
Seguimiento profesor	76.69	103.06*	2787.000	80.15	103.65*	2346.000
En conjunto, el sistema de EC	65.60	108.38***	2110.500	46.41	111.98***	1030.000
Valoración uso de:						
Clases	68.06	107.20***	2260,500	59.10	108.85***	1525,000
Actividades autónomas	63.88	109.21***	2005,500	58.23	108.50***	1491,000
Trabajo dirigido	89.32	96.99	3557,500	81.51	103.32*	2399,000
Valoración general:						
Estructuración de la asignatura	68.34	106.12***	2270,500	55.24	108.35***	1358,000
Actividades complementarias CV	72.99	104.83***	2561,500	71.58	105.77*	2011,500
Apuntes de clase	55.65	113.16***	1503,500	61.83	108.17***	1631,500
TIC en docencia	76.69	101.70*	2787,000	68.84	105.06***	2564,000
CV interacción profesorado-alumnado	83.42	99.12	3197,500	55.81	108.55***	1396,500
CV interacción entre iguales	83.20	99.93*	3184,000	62.62	107.98***	1662,000
CV ayuda en el aprendizaje	70.64	105.96***	2418,000	52.00	110.60***	1248,000
CV ayuda en motivación	79.12	101.89*	2935,500	58.27	109.05***	1492,500

Figura 4. Elementos de la innovación más relevantes

Mediante la prueba *U de Mann-Whitney* se obtienen las relaciones de los ítems del EMID adaptado con la alta y baja COMPRENSIÓN y SATISFACCIÓN:

Concordancias: se observa que las clases y los «problemas campus» son un elemento valorado tanto para los que afirman tener una alta comprensión de la

asignatura como los que manifiestan una alta satisfacción con el modelo propuesto. Además, la estructura de la asignatura (teoría, prácticas y trabajo dirigido), los apuntes de teoría y el uso del campus virtual como ayuda para el aprendizaje parecen ser factores clave para la comprensión y la satisfacción.

Discrepancias: en el trabajo dirigido y en el seguimiento solo parecen observarse leves diferencias a nivel de satisfacción en el primero y en ambos en el segundo. Las actividades complementarias parecen haber incidido más en la comprensión y, en cambio, las TIC y la interacción mediada, así como la motivación del campus, parecen estar más vinculadas a las puntuaciones altas con la satisfacción.

Por último, «en conjunto el sistema de evaluación continua me parece adecuado» obtiene en ambas variables los rangos promedio mayores puntuaciones altas.

CONCLUSIONES

Se ha evaluado un modelo de innovación docente de una asignatura metodológica (DEIA) adaptada a las directrices del EEES para ser implementada en el Grado de Psicología. Se basa en la evaluación continua, mediada por un campus virtual diseñado entorno a las actividades del estudiante y gestionado por el equipo docente. Los estudiantes han manifestado su positiva valoración del modelo europeo, señalando aspectos clave. En la actualidad se ha iniciado el grado en Psicología con los planes docentes europeos, la evaluación continua, clases magistrales de teoría y prácticas desdobladas, trabajando en equipos docentes y utilizando el campus virtual Moodle.

NOTA DE LOS AUTORES

Agradecimiento al proyecto de innovación docente REDICE08 de la Universitat de Barcelona.

REFERENCIAS

Bono, R., Arnau, J. y Blanca, Mª.J. (2005). Tecnologías de la información y comunicación en la enseñanza de diseños experimentales y aplicados. *Psicothema, 18*(3), 646-651.

López, O., Viader, M., Cosculluela, A., Malapeira, J.M., Rifà, X. y Cifre, I. (2008). Encuesta de innovación docente universitaria en Diseños experimentales y aplicados implementada en un campus virtual moodle. Ponencia presentada en el *Congreso de Metodología de Encuestas*, 2008, septiembre, Córdoba, España: IESA y Universidad de Córdoba.

López, O., Viader, M., Cosculluela, A., Honrubia, M.L. y Malapeira, J.M. (2009). Diseños experimentales y aplicados: innovación en evaluación continuada

basada en el feedback evaluativo del aprendizaje autónomo a través del CAMPUS VIRTUAL y el sistema de pruebas de validación. Ponencia presentada en el *VI Congreso Internacional de Docencia e Innovación Universitaria*, Barcelona, Julio 2010.

López-Fernández, O., Viader Junyent, M., Cosculluela Mas, A., Honrubia Serrano, M.L., Malapeira Gas, J.M.; Pirla Buil, L.; Aparicio Lopez, N. y Manzano Diaz, L. (en prensa). Perspectiva de los estudiantes respecto a la adaptación de «Diseños Experimentales y Aplicados» al Espacio Europeo de Educación Superior. *Revista d'Innovació i Recerca en Educació (REIRE)*.

Martínez, M. y Viader, M. (2008). Reflexiones sobre aprendizaje y docencia en el actual contexto universitario. La promoción de equipos docentes. *Revista de Educación* (número extraordinario), 213-234.

Universitat De Barcelona (2006a). Normes reguladores de l'avaluació i la qualificació dels aprenentatges. Barcelona: Publicacions i Edicions.

Universitat De Barcelona (2006b). Normes reguladores dels plans docents de les assignatures per als ensenyaments de la Universitat de Barcelona segons les directrius de l'Espai Europeu d'Educació Superior. Barcelona: Publicacions i Edicions.

Universitat De Barcelona (2007). Proyecto Piloto del Campus virtual de la UB. Universitat de Barcelona: Area de tecnologies – Equip Campus.

Viader, M., López, O., Rifà, X., Cifré, I., Cosculluela, A. y Malapeira, J.M. (2007). Incorporació de la plataforma Moodle com a eina de suport a la docència del marc europeu en l'assignatura Dissenys experimentals i aplicats. Ponencia presentada en la *Quarta Trobada de Professorat de Ciències de la Salud*, 2007, junio, Barcelona, España: Universitat de Barcelona.

APLICACIÓN Y EVALUACIÓN DE LAS DESTREZAS ADQUIRIDAS EN EL PROGRAMA DE FORMACIÓN DEL PROFESORADO NOVEL –MODALIDAD CONSOLIDACIÓN– PARA LA MEJORA DE LA ACTUACIÓN DOCENTE

José Manuel Sevillano Armenta, Milagrosa Sánchez Martín, Susana Sanduvete Chaves y Salvador Chacón Moscoso
Universidad de Sevilla
Correo electrónico: josesevillano@us.es

Resumen

Este trabajo supone la continuación de la experiencia llevada a cabo a través de la participación en las actividades de Formación de Profesorado Novel de la Universidad de Sevilla en la modalidad de consolidación. Haciendo uso de estrategias de investigación aplicada establecimos como objetivo general mejorar la comprensión de los conceptos estadísticos básicos por parte del alumnado e ir progresando hacia el desarrollo de una metodología docente basada en proyectos en dos de las asignaturas en las que se impartía docencia durante el curso 2010/2011. Para su implementación se hizo uso de un checklist de valoración de la docencia, reuniones de equipo para la programación de contenidos y ejercicios a desarrollar, bibliografía centrada en la impartición de la estadística a nivel universitario, diverso software y material estadístico y entrevistas en profundidad. Los resultados ofrecieron una notable mejora de las puntuaciones medias en la evaluación de las asignaturas consideradas, así como una elevada satisfacción con la ejecución del profesorado y el sistema de evaluación utilizado por parte del alumnado, junto a una mejora de la actuación docente como consecuencia de la propia intervención.

Este trabajo ha supuesto la continuación de los desempeños llevados a cabo durante el Programa de Formación del Profesorado Novel (PFPN) en su modalidad iniciación durante el curso académico 2009/10. Durante dicho programa, el objetivo del grupo de trabajo fue mejorar la actividad docente del profesorado novel mediante la participación activa en el PFPN. Para ello, se elaboró un Listado de Verificación de la Actuación Docente (LVAD), instrumento compuesto por 35 indicadores estructurados en 5 áreas de intervención, con la intención de facilitar la valoración de la actuación docente del profesorado señalando puntos fuertes y débiles; así mismo, se llevaron a cabo actividades específicas que finalmente resultaron en una mejora de la actuación docente. A estos desarrollos, se ha dado difusión a través: (a) un artículo publicado en la Revista de Enseñanza Universitaria (Sánchez-Martín, Sanduvete, Sevillano, Muñoz-Fernández y Chacón, 2010); y (b)

un capítulo de libro titulado «Programa de Formación del Profesorado Novel de la Universidad de Sevilla; Edición 2009/10. Análisis de su evaluabilidad» (Sánchez-Martín, Sanduvete y Chacón, 2011).

Como continuación, el PFPN en su modalidad consolidación supuso la profundización en conocimientos, habilidades y problemas concernientes a la práctica docente habitual de los profesores noveles; concretamente, se pretendía realizar una primera aproximación a un problema detectado en el proceso de enseñanza-aprendizaje de la estadística.

El grupo de trabajo quedo conformado por profesores del Departamento de Psicología Experimental (Área de Metodología de las Ciencias del Comportamiento), donde imparten asignaturas con contenidos de diseños de investigación a partir del método científico, que requiere un alto componente de análisis de datos.

En base a una estrategia de investigación aplicada (Varas y Rubio, 2004), se llevó a cabo como paso previo a la intervención en sí misma, una evaluación de necesidades con objeto de detectar en el alumnado aquellas limitaciones que se considerasen de mayor relevancia. Este proceso, nos ayudó a detectar la problemática existente en torno a la enseñanza de la estadística, evidenciando que el alumnado novel universitario muestra una problemática respecto a las asignaturas de estadística que gira en tornos a cuatro factores fundamentales: el profesorado, la enseñanza previa, las características de la materia y las características del alumnado.

LA FORMACIÓN Y MOTIVACIÓN DE LOS PROFESORES

La formación didáctica de los profesores debe incluir no solo conocimientos estadísticos sino el «conocimiento didáctico del contenido» (Batanero, 2001a; Batanero, Garfield, Ottaviani y Truran, 2000; Batanero, Godino y Roa, 2004), cuyos componentes básicos son (Batanero, 2002) : (a) la reflexión epistemológica sobre el significado de los conceptos, y procedimientos que se pretende enseñar, su desarrollo y evolución; (b) el análisis de las transformaciones del conocimiento para adaptarlos a los distintos niveles de enseñanza; (c) el estudio de las dificultades, errores y obstáculos de los alumnos en el aprendizaje y sus estrategias en la resolución de problemas; (d) análisis del currículo, situaciones didácticas, metodología de enseñanza para temas específicos y recursos didácticos específicos.

INCORPORACIÓN DE LA ESTADÍSTICA DESDE LA ESCUELA

Se detecta una escasa base de los alumnos en las materias de estadística, debido a la falta de incorporación de la estadística en la escuela o el éxito logrado en ello (Batanero, 2002). Esto conlleva que: (a) el profesorado universitario tenga que acelerar las explicaciones, suprimir actividades prácticas y demostraciones o razonamientos de gran utilidad para la comprensión del alumno; (b) al alumno le cuesta asimilar el contenido en un tiempo tan limitado).

CARACTERÍSTICAS DE LA MATERIA

Las asignaturas de estadística se caracteriza por (Batanero, 2000; Batanero y Godino, 2005; Batanero et al., 2000; Díaz, Batanero y Wilhelmi, 2008; Wulff y Wulff, 2004): (a) presentar multitud de conceptos abstractos; (b) los cuales requieren de conocimientos previos para poder asimilarse comprensivamente; (c) se basan, sobre todo, en la comprensión y resolución de problemas prácticos; (d) los cual requiere la realización de un gran número de casos prácticos; (e) al ser conceptos jerárquicamente relacionados precisan de un trabajo continuado a lo largo del curso por parte del alumnado.

CARACTERÍSTICAS DEL ALUMNADO

Finalmente, tampoco se debe olvidar que existen alumnos con diferente predisposición para la materia; diferentes etnias (culturas); edades, sexo, orientación sexual, estatus socioeconómico, etc., así como diferentes estilos de aprendizaje (Harris, Mazoué, Hamdan y Casiple, 2007; Lesser, 2010; Stephenson, 2010). Estas características particulares de los alumnos los harán más o menos predispuestos a asimilar los conocimientos estadísticos.

Como consecuencia de la conjunción de estos factores, suelen detectarse dificultades en los alumnos. Concretamente: (a) suelen darse problemas debido a factores actitudinales o creencias negativas sobre la estadística. Existe generalmente un escaso interés en la materia, llegando a percibir la estadística como algo difícil y desagradable (Wulff y wulff, 2004); (b) no dominan el cálculo y son pocos los que controlan las ideas básicas de probabilidad (Batanero, et al. 2000; Batanero, Godino, Green y Vallecitos, 1994); (c) no comprenden por qué la estadística se incluye en sus curricula; no ven cómo la estadística les puede ayudar (Cordani, 2000).

En este contexto, con objeto de abordar la problemática detectada, las teorías de aprendizaje enfatizan diversos aspectos de gran relevancia en la construcción del conocimiento (Batanero, 2002), como son el papel de la resolución de problemas, la formulación (lenguaje matemático), la validación (demostración y razonamiento de las ideas matemáticas) y la institucionalización (puesta en común; acuerdo social).

Así, cobran un papel primordial los ejercicios adaptados, los proyectos estadísticos y la experimentación (Batanero, 2002; Batanero, n.d.; Cordani, 2000), ya que ayudan a los estudiantes a aprender estadística y al mismo tiempo incrementan sus capacidades de innovación, creatividad y su actitud crítica.

En este sentido, el objetivo que propuso el grupo de trabajo fue mejorar la comprensión de los conceptos estadísticos básicos por parte del alumnado, lo que conllevaría resultados finales más positivos. Sin embargo, al ser un objetivo demasiado ambicioso, se planteó un procedimiento que fuera el primer paso para ir avanzando hacia el desarrollo de «proyectos» (los cuales se pretenden implantarán

en próximos cursos), que previsiblemente potenciarán la motivación del alumnado hacia el estudio de la estadística y la comprensión de los conceptos involucrados.

MÉTODO

Participantes

Se distinguieron tres conjuntos diferentes de participantes:

A. Componentes del grupo.

El grupo estuvo compuesto de un profesor mentor y tres noveles pertenecientes al Área de Metodología de las Ciencias del Comportamiento (Departamento de Psicología Experimental): (a) El profesor mentor fue un varón de 43 años con 19 años de experiencia en docencia e investigador principal de un grupo de investigación; (b) los tres profesores noveles que finalizaron el programa fueron dos mujeres de 27 y 30 años y un varón de 30, todos ellos con menos de cinco años de experiencia docente universitaria.

B. Alumnado participante en el proceso de detección de necesidades.

Fueron 44 estudiantes de la asignatura «Metodología Observacional» y 17 estudiantes de la asignatura «Diseño y Análisis de Datos en Psicología I».

C. Alumnado participante en la intervención.

Se trató de un total de 217 matriculados en los grupos donde impartió docencia el profesorado novel participante y tuvo acceso a los resultados finales; concretamente, 93 matriculados en «Metodología observacional» en el curso 2010/11; y 124 matriculados en la misma asignatura en el curso 2009/10.

Instrumentos

Los instrumentos utilizados para el desarrollo de este trabajo fueron los siguientes:

A. Listado de Verificación de la Actuación Docente (LVAD).

Para llevar a cabo la detección de necesidades, se utilizó un «checklist» para la evaluación de la actuación docente del profesorado novel, elaborado por el grupo de profesores durante el PFPN en su modalidad iniciación. El instrumento se compone de 35 indicadores para la mejora de la actuación docente, estructurados en 5 áreas diferentes. Cada indicador se valora haciendo uso de una escala de tres puntos (asociados a consecución baja, media y alta).

B. Bibliografía.

Para la elaboración de los materiales se utilizó la documentación disponible en la plataforma virtual de la Universidad de Sevilla, en la sección habilitada para el

PFPN (fundamentalmente de los módulos Aprendizaje y Metodología); así mismo, se utilizó diversa documentación relacionada con la enseñanza de la estadística (la cual se encuentra referenciada en su mayoría en la introducción de este manuscrito, así como en las conclusiones y desarrollos futuros).

C. Software

Se hizo uso de Excel para la codificación de datos y análisis descriptivos, y el paquete estadístico SPSS v17.0 para los contrastes de hipótesis.

D. Entrevista en profundidad

Se elaboró un protocolo de entrevista a partir de las dimensiones evaluadas en el LVAD (Sánchez-Martín, Sanduvete, Sevillano, Muñoz y Chacón, 2010) y sirvió como guión para la realización de entrevistas en profundidad al alumnado.

Procedimiento

A. Detección de necesidades

En primer lugar se administró el LVAD a la muestra de estudiantes. Con objeto de obtener información adicional sobre la evaluación, se añadió la pregunta «¿crees que la evaluación se ajusta a los contenidos impartidos?» al final del examen del primer parcial de la asignatura «Metodología Observacional».

B. Búsqueda de bibliografía complementaria

Con objeto de obtener más información sobre la didáctica de la estadística, sus características y problemática relacionada se llevó a cabo una búsqueda bibliográfica exploratoria en las bases de datos de interés disponibles en la Universidad de Sevilla (EBSCO Online, Medline, Serfila, CABHealth, CINAHL, Econlit, MathSci, Current Contents, Humanities Index, ERIC and PsycINFO), combinando las palabras clave «estadística», «didáctica» y «profesorado».

C. Elaboración de materiales adaptados

En base a la detección de necesidades previas y al sustento bibliográfico se diseñaron y realizaron ejercicios prácticos de cada tema durante las clases, intentando no solo que incluyera la materia pertinente sino que estuvieran elaborados de forma que el alumno pudiera ver su utilidad práctica o les resultara de temática interesante. Una parte de estos ejercicios los debía resolver el alumno con el asesoramiento del profesor (Moore, 1997), mientras que otra parte debía resolverla individualmente para la consolidación del aprendizaje y el fomento de la autonomía del alumno.

D. Ejercicios de grupo pequeño.

En las clases prácticas se llevó a cabo la resolución de actividades en grupos de trabajo de forma que los estudiantes construyesen su conocimiento de forma acti-

va, mediante la resolución de problemas y la interacción con sus compañeros (Moore, 1997).

E. Reuniones de grupo.

Se llevaron a cabo varias reuniones de grupo. La primera para planificar el trabajo y el proceso de detección de necesidades. La segunda para analizar los resultados obtenidos de la detección de necesidades y delimitar el proceso de búsqueda de bibliografía relacionada, así como el diseño de la intervención y el posterior análisis de resultados. La tercera con objeto de poner en común la información obtenida y comenzar a revisar y modificar los materiales. Tres reuniones más se dedicaron a continuar adaptando los materiales. Finalmente la séptima reunión se desarrolló para evaluar los resultados obtenidos.

Resultados

A. Resultado de las evaluaciones

La evaluación final de la asignatura de Metodología observacional registró un importante cambio en los cursos 2009/2010 y 2010/2011. Así, y atendiendo al porcentaje de aprobados se observa una notable mejora durante este último año (94,7% frente a un 77,2%; χ^2 (1)= 10,078**; phi=0,239**). A nivel cuantitativo supuso, además, un incremento significativo en las puntuaciones medias de la asignatura a favor del último curso académico (7,45 frente a 5,86; F(1,174)= 32,832**), obteniéndose un tamaño de efecto medio (eta cuadrado parcial = .159).

Esta tendencia también se observa cuando se consideran los datos de forma ordinal, encontrándose una mayor cuantía de sujetos con calificaciones altas (notable, sobresaliente y matrícula de honor) durante el último año.

B. Puntuaciones en el LVAD

Tras analizar las respuestas del alumnado en el LVAD, se observó una alta satisfacción con las características de la docencia según las dimensiones consideradas, puesto que en las dos asignaturas valoradas (Análisis de Datos en Psicología I y Metodología Observacional) más del 89% de los indicadores superaban la puntuación 2 (valor medio de la escala cuyo rango fue 1-3).

En concreto, aquellos ítems que ofrecieron una mejor valoración fueron: *Evaluación continua, seguimiento a lo largo de todo el curso* (Media = 2,82) y *Explicación clara del contenido teórico* (Media = 2,78) para la asignatura de Metodología observacional (n=44), y *Mantenimiento del aula en silencio durante las explicaciones* (Media = 2,88), *Logro de buen clima* (Media = 2,88) y *Motivación al alumnado para que participe y explicite sus dudas* (Media = 2,88) para Análisis de Datos en Psicología I (n=17). Por otro lado, aquellos que registraron menor valoración fueron: *Planificación de trabajo en grupo* (Media = 1,97) y *Finalización de la clase con un resumen de todo lo visto en ella* (Media = 1,77) para la primera asignatura e *Inicio de la clase con un resumen de lo visto hasta el momen-*

to (Media = 1,47) y *Resolución clara de las actividades* (Media = 1,82) para la segunda.

C. Entrevistas en profundidad.

Al finalizar el cuatrimestre se solicitó la colaboración del alumnado de forma voluntaria para participar en una entrevista de profundidad (Rubio y Varas, 1997) que se desarrollaría de acuerdo a un protocolo elaborado a partir de los elementos más destacados de la actuación docente recogidas en la herramienta LVAD (Sanduvete, Sánchez, Sevillano, Muñoz y Chacón, 2010).

De este modo, en relación a la impartición de la asignatura, los entrevistados destacaron que el desarrollo de la misma no se adaptó de forma dinámica al grueso del alumnado, sino que su docencia se focalizó principalmente en aquellos grupos con menor competencia (*«Se quedó en el grupo de nivel muy bajo, en vez de tirar por el camino medio, se quedó atrás… no suprimía las cosas que supuestamente ya sabíamos; me pareció un poco repetitivo… nos aburríamos y no profundizaba en lo que de verdad es importante.»* [Entr. 1]), no realizándose una adaptación dinámica y acorde al nivel medio de los grupos. Sin embargo, su comunicación sí fue muy apreciada por la cercanía que se transmitía en las interacciones (*«Era como muy cercano. Más bien motiva que cohíbe. No te da vergüenza preguntarle»* [Entr. 1]. *«La ves muy cercana, te da confianza, la ves tan llana que no te da vergüenza preguntarle»* [Entr. 2]).

Por otro lado, también fueron muy valorados los nuevos ejercicios que se desarrollaron durante el proceso seguido en este curso académico (*«Sí, facilitan…lo importante es ver dónde puedes utilizarlo (aplicar los contenidos) en un futuro si te dedicas a eso.» «Ponerte en clase los ejercicios de examen…eso estaba muy bien; a mí me gustaban mucho.»* [Entr. 2]), ya que hacían hincapié en problemas habituales de la práctica psicológica y se veía de una forma más clara su aplicación profesional. No obstante, y aunque durante la elaboración del listado de valoración de la actuación docente se reconoció la aportación que suponía la presentación de un resumen tanto al inicio como al final de la clase, se observó que ambos resúmenes tuvieron una aceptación diferente. Así, el resumen del inicio sí era valorado positivamente, puesto que situaba al alumno al inicio de la clase; sin embargo, el resumen final supuso una carga adicional a la atención que no repercutió en mejoras del aprendizaje, por lo que sería conveniente cuestionar su aplicación e inclusión en dicha herramienta (*«Las clases son muy largas, dos horas sin parar. Te quieres ir y te da igual el resumen. Normalmente comienza recopilando, pero el resumen final… (suspiro)… al inicio te sitúa»* [Entr. 1]).

En cuanto a la utilización de software informático (mediante presentaciones en powerpoint principalmente) se destacó la utilidad que ofrecía para centrarse en elementos críticos de la asignatura («*Cuando lo leo en las diapositivas sé donde hacer más hincapié»* [Entr. 2]), así como herramienta de síntesis de los contenidos («*… es como un esquema, un resumen del tocho gigante de la asignatura»* [Entr. 1]).

Finalmente el sistema de evaluación continua supuso a los alumnos mayores ventajas que la evaluación sumativa final pues según estos resultaba *«…cómoda para aprobar y entender la asignatura… tienes más posibilidades, pues puedes cambiar de método de estudio si uno no te sirve* [Entr. 2], aunque, no obstante, se percibió como insuficiente la docencia impartida sobre aplicaciones de software estadístico *«añadiría más peso a las prácticas»* [Entr. 1]. Esta valoración coincidía con las respuestas ofrecidas a la pregunta «*Crees que la evaluación se ajusta a los contenidos impartidos*» donde el 76% del alumnado (n=72) estaba muy de acuerdo con el sistema de evaluación y solo un escaso 0,04% no estuvo de acuerdo.

DISCUSIÓN Y CONCLUSIONES

Consideramos que el desarrollo del PFPN en su modalidad de consolidación nos ha sido de gran ayuda, como profesores de estadística aplicada a la psicología, para comprender mejor las dificultades que presenta un gran porcentaje de nuestro alumnado a la hora de comprender los conceptos de forma significativa, y como enfocar la didáctica de estas materias para potenciar su aprendizaje.

Respecto a los resultados obtenidos tras la aplicación de una metodología basada en ejercicios adaptados, se ha obtenido: (a) una mejora de los resultados en las asignaturas objeto de intervención con respecto al curso anterior; (b) mayor conocimiento y aprendizaje del alumnado; (c) alta satisfacción del alumnado medida a través del cuestionario LVAD (89% de los indicadores ≥2 puntos); (d) mejora de la actuación docente como consecuencia de la propia intervención.

Esta intervención constituye un paso intermedio para poder instaurar una metodología basada en proyectos en los próximos cursos académicos (Batanero, 2000). Además, nos gustaría llevar a cabo otros desarrollos futuros, como: a) incorporar activamente las nuevas tecnologías a las clases de estadística para hacerlas más atractivas y entretenidas para el alumnado (de Sousa y Novegil, 2010); b) desarrollar ayudas interactivas; c) administrar a los alumnos un cuestionario previo para analizar sus tipos de inteligencia y diseñar actividades en función de dichos resultados (Piper, 2002); d) llevar a cabo el proceso de validación de LVAD, con objeto de que llegue a ser una herramienta útil y fiable para detectar los puntos fuertes y débiles de la actuación docente.

REFERENCIAS

Batanero, C. (2000). ¿Hacia dónde va la educación estadística? *Blaix, 15*, 2-13.

Batanero, C. (2001), *Didáctica de la estadística*. Granada: La autora.

Batanero, C. (2002). Los retos de la cultura estadística. Conferencia inaugural de las *Jornadas Interamericanas de Enseñanza de la Estadística*. Buenos Aires.

Batanero, C. (n.d.). ¿Por qué y cómo enseñar Estadística? Disponible en http://www.ugr.es/~batanero/

Batanero, C., Garfield, J. B., Ottaviani, M. G. y Truran, J. (2000). Investigación en Educación estadística: Algunas Cuestiones Prioritarias. *Statistical Education Research Newsletter, 1*(2).

Batanero, J. C., Godino, D. R. Green, P. Colmes y Vallecitos, A. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology, 25*(4), 527-547.

Batanero, C., Godino, J. D. y Roa, R. (2004). Training teachers to teach probability. *Journal of Statistics Education, 12*(1).

Cordani (2000). Comentario sobre el artículo Investigación en Educación estadística: Algunas Cuestiones Prioritarias. *Statistical Education Research Newsletter, 1*(2), 10-11.

De Sousa, B. y Souto, J. V. (2010). How technology can help or complicate the teaching of statistics depending on the class size. En C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. Eighth International Conference on Teaching Statistics (ICOTS8, Julio, 2010), Ljubljana, Slovenia.

Diaz, C., Batanero, C. y Wilhelmi, M. (2008). Errores frecuentes en el análisis de datos en educación y psicología. *Publicaciones, 35,* 109-123.

Harris, C. M., Mazoué, J. G., Hamdan, H., y Casiple, A. R. (2007). Designing An Online Introductory Statistics Course. En D. S. Dunn, R. A. Smith y B. C. Beins (Eds.), *Best Practices for Teaching Statistics and Research Methods in the Behavioral Sciences* (pp. 93-108). London: Lawrence Erlbaum Associates.

Lesser, L. M. (2010). En C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. Eighth International Conference on Teaching Statistics (ICOTS8, Julio, 2010), Ljubljana, Slovenia.

Moore, D. S. (1997). New Pedagogy and new content: the case of statistics. *International Statistical Review, 65*(2), 123-165.

Piper, C. (2002). *Multiple Intelligence Quiz*. Disponible en: http://www1.chapman.edu/soe/faculty/piper/teachtech/miquiz.htm.

Sánchez-Martín, M. Sanduvete, S. Sevillano, J. M., Muñoz, N. y Chacón, S. (2010). Reflexiones sobre la participación y adquisición de destrezas para la mejora de la actuación docente en el programa de formación del profesorado novel. *Revista de Enseñanza Universitaria, 35,* 41-53.

Sánchez-Martín, M., Sanduvete, S. y Chacón, S. (2011). Programa de Formación del Profesorado Novel de la Universidad de Sevilla; Edición 2009/10. Análisis de su evaluabilidad. En A. Aguilera Jiménez y M. Gómez de Terreros Guardio-

la (Coords.). *Actividad docente en el marco del espacio europeo de educación superior* (pp. 306-317). Sevilla: Fenix Editora.

Stephenson, W. R. (2010). Diversity and differentiated instruction and learning. En C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. Eighth International Conference on Teaching Statistics (ICOTS8, Julio, 2010), Ljubljana, Slovenia. Disponible en www.stat.auckland.ac.nz/~iase/publications.php

Varas, J. y Rubio, M. J. (2004). *El análisis de la realidad en la intervención social: métodos y técnicas de investigación*. Madrid: CCS

Wulff, S. S. y Wulff, D. H. (2004), Scholarship of Teaching and Learning. «Of Course I'm Communicating; I Lecture Every Day»: Enhancing Teaching and Learning in Introductory Statistics. *Communication Education, 53*(1), 92-103.

METODOLOGÍA ABP (APRENDIZAJE BASADO EN PROBLEMAS) PARA LA DOCENCIA DE ANÁLISIS DE DATOS Y DISEÑOS DE INVESTIGACIÓN EN PSICOLOGÍA

Laura Vozmediano, Nerea Lertxundi, Ana Isabel Vergara, Arantxa Gorostiaga y Xabier Isasi
Universidad del País Vasco
Correo electrónico: laura.vozmediano@ehu.es

Resumen

En este trabajo se presenta el diseño de la asignatura «Análisis de datos y diseños: método no experimental» del Grado en Psicología para su impartición con metodología ABP (Aprendizaje Basado en Problemas) y técnicas de aprendizaje cooperativo. Esta propuesta se enmarca en el programa ERAGIN para la formación del profesorado en metodologías activas de enseñanza, que la Universidad del País Vasco / Euskal Herriko Unibertsitatea desarrolla como parte de su Modelo IKD de enseñanza-aprendizaje cooperativo y dinámico. Este programa pretende desarrollar competencias docentes específicas que se orientan a conseguir que el estudiante participe activamente en su aprendizaje. La metodología ABP nos ofrece herramientas para promover un aprendizaje significativo y activo. Se emplea como punto de partida un problema/escenario cuya resolución requiere alcanzar ciertos resultados de aprendizaje –a través del trabajo cooperativo– permitiendo así adquirir las competencias de la asignatura. El trabajo presenta la labor desarrollada por el equipo docente en el diseño de objetivos de aprendizaje coherentes con las competencias a adquirir, en la propuesta de las actividades y técnicas a emplear en cada sesión con el alumnado, así como en la concreción de los resultados de aprendizaje esperados junto a su correspondiente procedimiento de evaluación.

La Universidad del País Vasco / Euskal Herriko Unibertsitatea se interesa desde hace años por la innovación educativa, y en particular por las nuevas metodologías docentes, impulsando la formación docente y la puesta en marcha de experiencias innovadoras (por ejemplo, Garaizar y Goñi, 2010). Una muestra de ello es el programa ERAGIN para la formación del profesorado en metodologías activas de enseñanza, que se enmarca en el Modelo IKD (Ikasketa Kooperatiboa eta Dinamikoa) de enseñanza-aprendizaje cooperativo y dinámico de esta Universidad. El programa pretende desarrollar competencias docentes específicas que se orientan a conseguir que el estudiante participe activamente en su aprendizaje. Esto es especialmente relevante en un contexto económico, social y cultural cambiante, en el que los futuros profesionales precisan de competencias tales como la capacidad de

organizar y analizar críticamente grandes cantidades de información, el dominio de las tecnologías de información y comunicación, y la capacidad para trabajar en equipo y dirigir sus procesos de aprendizaje.

El programa ofrece tres alternativas de metodologías activas, habiéndose escogido el ABP para diseñar e impartir una parte de la asignatura Análisis de datos y diseños: método no experimental (Troncal de 2º curso/ 6 ECTS). ABP es una metodología activa que fomenta el aprendizaje como proceso constructivo y no receptivo, el aprendizaje autodirigido, y el aprendizaje significativo y contextualizado. Se emplean técnicas que posibilitan que el alumnado establezca sus propias metas de aprendizaje, en el proceso de resolución de un problema/escenario significativo. Fijadas las metas, el estudiante tendrá que adquirir los conocimientos y habilidades concretas, y ponerlas en marcha para resolver finalmente el problema. El programa ERAGIN integra, además, técnicas de aprendizaje cooperativo, por lo que gran parte de estas tareas se realizarán en equipo, de modo que el éxito individual dependerá del trabajo realizado por todos los componentes del equipo.

Los contenidos del temario que se diseñaron con ABP fueron los siguientes:

- Bloque 1. El proceso de la investigación científica. Revisión de la literatura y planteamiento de objetivos e hipótesis. Diseño de un estudio con método no experimental (muestra, materiales y procedimiento).
- Bloque 2. Software para la investigación II. Análisis de supuestos del modelo estadístico mediante SPSS. Análisis de potencia mediante G-power. Cálculo del tamaño del efecto con Excel. Cálculo del tamaño muestral con Excel. Descriptivos con SPSS. Análisis de correlación y tablas de contingencia mediante SPSS. Regresión lineal mediante SPSS.

En este trabajo se presenta el proceso de formación del profesorado y los resultados de la primera fase del programa: el diseño de la asignatura y realización de los materiales para su posterior implementación, en una segunda fase que todavía están en curso.

PROCESO DE FORMACIÓN DEL PROFESORADO

La formación del programa ERAGIN comenzó con un taller de iniciación en el que se contó con la participación de expertos internacionales en ABP como Luis Branda, que presentó las características principales de esta metodología docente (Branda, 2009) y aportó ejemplos de su dilatada experiencia trabajando con ABP. Asimismo participó profesorado formado en anteriores ediciones del programa; algunos de estos profesores ya formados pasaron a ser tutores de quienes realizaban la formación en esta edición.

Durante 6 meses se desarrolló la labor de diseño y elaboración de materiales, contando con la supervisión del tutor. De hecho, la figura de tutor es uno de los aspectos clave del programa, ofreciendo un punto de apoyo que facilita la labor de

los docentes que se están formando, al tiempo que realiza una primera evaluación interna de los materiales, que garantiza un nivel de calidad suficiente para la posterior evaluación externa. El diseño evaluado positivamente por el tutor pone fin a la primera fase del programa.

La segunda fase, que no se describe en este trabajo por encontrarse todavía en curso, comprende la implementación del diseño y el informe final que será evaluado externamente para la superación del programa.

Las tareas que los docentes han acometido en esta primera fase, de modo sintético, son las siguientes:

- Diseño del Curriculum. Incluye la descripción del contexto de la asignatura, definición de competencias y resultados de aprendizaje, diseño y secuenciación de actividades, elaboración de materiales…
- Trabajo en un proceso colaborativo, que cuenta con el asesoramiento del tutor/a, en el que se comparten materiales, dudas e inquietudes en foros colaborativos, y en el que se realizan reuniones presenciales y no presenciales.
- Obtención de retroalimentación. Se trabaja con un sistema de entregables sucesivos, obteniendo feedback del tutor en cada entrega, de modo que el siguiente entregable incorpore las sugerencias de mejora del anterior.

RESULTADOS DE LA 1ª FASE: DISEÑO

Uno de los resultados principales de la labor desarrollada es la concreción de las competencias de asignatura, curso y grado que se trabajarían en los bloques a impartir con ABP (ver figura 1). Esta tarea forma parte de un proceso de reflexión más amplio, que desemboca finalmente en el diseño de un «problema estructurante» que se presentará al alumnado. Ése es el escenario que habrá de resolver, a través de una serie de tareas sucesivas. Para conseguirlo, ha de alcanzar los objetivos de aprendizaje, que se corresponden con el temario a impartir, y que garantizan que una vez cumplidos, han contribuido a adquirir las competencias. La coherencia entre estos elementos es la clave del diseño.

Para el bloque 1, los objetivos de aprendizaje fueron:

- Realizar búsquedas de literatura científica en las fuentes de datos relevantes para las Ciencias Sociales y de la Salud.
- Plantear objetivos relevantes e hipótesis de investigación coherentes que puedan alcanzarse a través de estudios con método no experimental.
- Definir la población objeto de estudio, el tamaño muestral y tipo de muestreo a emplear.
- Escoger materiales adecuados para operacionalizar las variables de la investigación.

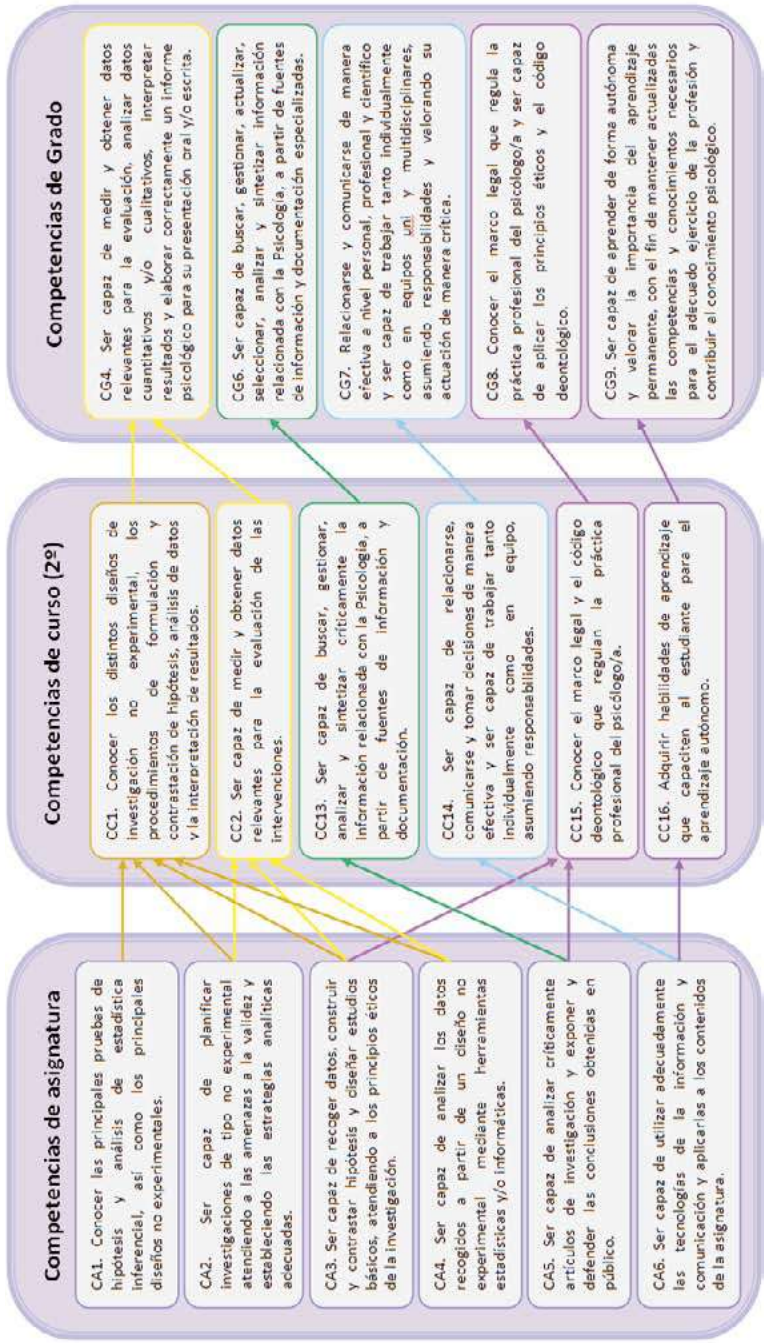


Figura 1 Competencias de asignatura, curso y grado que se trabajan en los bloques de la asignatura impartidos con ABP

– Escoger el procedimiento en un estudio con método no experimental.

Tales objetivos contribuían a las competencias siguientes:

– Ser capaz de utilizar adecuadamente las tecnologías de la información y comunicación y aplicarlas a los contenidos de la asignatura.

– Ser capaz de analizar críticamente artículos de investigación y exponer y defender las conclusiones obtenidas en público.

– Ser capaz de recoger datos, construir y contrastar hipótesis y diseñar estudios básicos, atendiendo a los principios éticos de la investigación.

– Ser capaz de planificar investigaciones de tipo no experimental atendiendo a las amenazas a la validez y estableciendo las estrategias analíticas adecuadas.

Respecto al bloque 2, los objetivos de aprendizaje fueron:

– Identificar el análisis estadístico y la herramienta estadística y/o informática que se precisa para dar respuesta a una pregunta de investigación en un estudio con método no experimental.

– Llevar a cabo el análisis estadístico e interpretar el resultado obtenido.

Tales objetivos contribuyen a la competencia:

– Ser capaz de analizar los datos recogidos a partir de un diseño no experimental mediante herramientas estadísticas y/o informáticas.

Además, hay dos objetivos de aprendizaje trasversales a ambos bloques de contenidos, que contribuyen igualmente a las competencias de asignatura, pero muy especialmente a las competencias trasversales del Grado en Psicología.

Son los siguientes:

– Redactar en adecuado estilo científico, siguiendo las normas de la APA, y respetando los principios éticos de la investigación con seres humanos.

– Trabajar en equipo de modo colaborativo y eficaz.

El Problema Estructurante, cuya resolución implica haber adquirido las competencias ya descritas, se reproduce a continuación:

Cómo llegar a ser el «fichaje estrella» de los equipos de investigación

Acabas de graduarte en Psicología y te interesa la investigación en temas sociales, ya que quieres realizar estudios con impacto en la sociedad, que mejoren la vida de los ciudadanos. Tu objetivo es incorporarte como asistente de investigación en un centro puntero de investigación en Ciencias Sociales y de la Salud, especializado en estudios sobre calidad de vida empleando el método no experimental. Este centro, el CIReSSH (Centre for Innovative Research in Social Scien-

ces and Health), ofrece la posibilidad de realizar prácticas de tres meses de duración, que sirven a su vez como proceso de selección.

Los investigadores en prácticas se organizarán en equipos y llevarán a cabo diversas tareas de investigación para el centro, supervisadas por el investigador principal a cargo de las prácticas y el proceso de selección.

Uno de los objetivos de las prácticas será diseñar, redactar y exponer un proyecto de investigación en el tema propuesto por el CIReSSH. El mejor de los proyectos será presentado a una convocatoria para obtener una subvención. El centro considera que un asistente de investigación debe ser autónomo para realizar revisiones de la literatura y diseñar proyectos sencillos bajo la supervisión de un investigador más experimentado.

Por tanto, en los tres meses de las prácticas, realizaréis la revisión de la literatura empleando las bases de datos online clave en Psicología, estableceréis los objetivos e hipótesis del estudio y diseñaréis la investigación, redactando todos estos aspectos en un proyecto de investigación coherente y relevante, que expondréis al final de las prácticas. Si el proyecto es de calidad y cumple los criterios para tener éxito en una convocatoria y llevarse a cabo, podrá aportar nuevos conocimientos a las Ciencias Sociales y de la Salud, y valor a la sociedad.

Un segundo objetivo de las prácticas tiene relación con los análisis estadísticos. Un asistente de investigación que trabaje en CIReSSH ha de saber analizar los datos de un estudio e interpretarlos de modo autónomo. Aprovechando que en el centro se están llevando a cabo los análisis de datos de un estudio sobre calidad de vida en la Comunidad Autónoma del País Vasco, el investigador principal proporcionará datos para llevar a cabo los análisis.

Por tanto, tu equipo será responsable de analizar los datos que se le presenten, para lo que deberá escoger –razonándolo- la herramienta de software apropiada y determinar cuál es el análisis estadístico más adecuado. A partir de esta decisión, realizaréis los análisis e interpretaréis el resultado que os ofrece, redactando una serie de informes de resultados para entregar al investigador principal del estudio.

Los equipos que completen los dos grandes objetivos que se le presentan (diseñar el proyecto y realizar e interpretar los análisis de datos) habrán demostrado que tiene los conocimientos clave de los análisis de datos y los diseños con método no experimental. Y además, que tienen capacidad de trabajo en equipo así como capacidad para resolver con un grado importante de autonomía situaciones a las que debe hacer frente un asistente de investigación. Obtendrán por tanto una evaluación positiva de sus prácticas y una recomendación del centro para su currículum. Pero sólo un equipo, el que haya demostrado mejor rendimiento en todas las tareas, será escogido por el CIReSSH para conceder una beca de dos años a sus miembros e iniciar una carrera como investigadores en Ciencias Sociales y de la Salud dentro del centro. Si en tu equipo sois los mejores, conseguiréis ser el «fichaje estrella» de este año.

Finalmente, se diseñaron las actividades concretas que en un sistema de pasos, permitirán la resolución del problema, se planificó el cronograma de las sesiones, especificando la modalidad docente, tipo de actividad (presencial vs. no presencial; individual, grupal, parejas…) y una descripción de las tareas a realizar (ver ejemplo en figura 2). Se anexaron los materiales necesarios para el desempeño de las actividades.

Esta labor de diseño se tradujo en un entregable final de la primera fase, que incluyó el Cuaderno del Profesor/a (con todas las indicaciones para que otro docente pudieran impartir los contenidos descritos con metodología ABP) y el Cuaderno del Alumno/a, que facilitará al alumnado el seguimiento de la asignatura en la implementación.

EJEMPLO SEMANA 3 (Del 19 al 23 de septiembre)	
ACTIVIDAD B1.1 ¿QUÉ ES UN "BUEN" OBJETIVO PARA UN PROYECTO DE INVESTIGACIÓN?	
A realizar ANTES DE LA PRIMERA SESIÓN DE TALLER	**TAREA 1** **NO PRESENCIAL & INVIDIVUAL**: Lectura Tiempo estimado: 15 minutos
SESIÓN 1 TALLER	**TAREA 2** **PRESENCIAL & INDIVIDUAL:** Exposición de la profesora **TAREA 3** **PRESENCIAL & EN PAREJAS:** Tormenta de ideas: ¿Qué es un "buen" objetivo? **TAREA 4** **PRESENCIAL & TODO EL AULA:** Puesta en común de ideas y recapitulación

Figura 2. Ejemplo de actividad del bloque 1

Por último, se diseñó un sistema de evaluación que ha de incluirse en el cuaderno del alumnado, para que el alumno/a sepa cómo enfocar sus esfuerzos. Concretamente en nuestro diseño se establecieron unos requisitos previos sin los cuales los entregables no serán evaluados: (1) Entrega en fecha, (2) Limpieza y formato adecuado, y (3) Corrección ortográfica y gramatical.

Los criterios de evaluación por bloque fueron:

Bloque 1:

10% Entregable final (PROYECTO DE INVESTIGACIÓN)

3% Exposición del grupo defendiendo su proyecto

3% Auto y co-evaluación del funcionamiento del grupo

TOTAL 16% de la asignatura

Bloque 2:

5% entregables de la actividad 6 a la actividad 10

3% entregable de la actividad 11

2% Auto y co-evaluación del funcionamiento del grupo

TOTAL 10%

Además de estos criterios generales, el cuaderno del alumno/a incluyó una definición concreta del modo de evaluación de cada entregable. Por ejemplo, en el caso de los informes de resultados del Bloque 2, se valora si:

- Analiza el problema planteado y detecta el análisis estadístico y herramienta apropiada
- Lleva a cabo el análisis obteniendo el resultado en el formato solicitado (tabla, gráfica, valor del estadístico…)
- Redacta el informe de resultados en adecuado estilo científico – indicador del objetivo trasversal

VALORACIÓN DE LA 1ª FASE DEL PROGRAMA

A pesar de la dificultad para identificar un problema estructurante, finalmente se pudo diseñar un problema que responde a las competencias que se pretenden desarrollar y que se espera resulte motivante para el alumnado, a través de la creación de una situación competitiva.

El resultado obtenido es valorado muy positivamente por el equipo, toda vez que se ha superado con éxito la evaluación de esta primera fase, lo que permite pasar a la implementación con buenas expectativas. Con la utilización de la metodología ABP y el aprendizaje cooperativo se espera conseguir una mayor implicación y motivación del alumnado con respecto a metodologías más tradicionales. Del mismo modo, se pretende lograr un buen nivel de adquisición de las competencias, en una asignatura del área de metodología que, a priori, puede ser considerada «difícil» o resultar poco motivadora para una parte del alumnado.

REFERENCIAS

Branda, L. A. (2009). El aprendizaje basado en problemas. De herejía artificial a *res popularis*. *Educación Médica, 12*, 11-23.

Garaizar, J. y Goñi, J.M. (2010). *Nuevos escenarios para el aprendizaje en la Universidad: Propuestas de innovación educativa de la UPV/EHU*. Bilbao: Servicio Editorial de la Universidad del País Vasco.

APLICACIONES METODOLÓGICAS

APLICACIÓN DEL ANÁLISIS SEMIÓTICO AL ESTUDIO DE ESTRATEGIAS EN LOS JUICIOS DE ASOCIACIÓN

Gustavo R. Cañadas, Carmen Batanero, José M. Contreras y Pedro Arteaga
Universidad de Granada
Correo electrónico: grcanadas@ugr.es

Resumen

El estudio de los juicios de asociación tiene una gran tradición en psicología, donde se ha analizado la precisión en dichos juicios en función de las estrategias empleadas, las variables en la tarea y el efecto de las teorías previas. Este trabajo describe un análisis cualitativo exploratorio de las estrategias empleadas en juicios de asociación en tablas de contingencia. La muestra estuvo formada por un total de 62 alumnos en la Licenciatura en Psicología, pues la evaluación tenía una finalidad diagnóstica y los resultados se usarían para proponer acciones educativas adecuadas. Los estudiantes respondieron a un cuestionario de respuesta abierta formado por cuatro tablas, en cada una de las cuáles se incluían tres apartados: (a) juicio de asociación; (b) estimación del coeficiente de asociación; y (c) razonamiento detallado de la estrategia seguida. Las variables independientes consideradas fueron: dependencia estadística, intensidad de la dependencia, concordancia entre los datos y las teorías previas, y tipo de covariación. Las respuestas obtenidas se analizaron en dos pasos: el análisis de contenido (Krippendorff, 1991), y usando elementos teóricos del enfoque ontosemiótico (Godino, Batanero y Font, 2007). Los resultados muestran una alta proporción de estrategias parcialmente correctas o incorrectas, así como numerosos conflictos semióticos.

Una de las formas más comunes de presentar la información estadística es en forma de tabla de doble entrada o tabla de contingencia (ver ejemplo en Tabla 1), a las cuales se presta poca atención en la enseñanza universitaria, suponiendo que su lectura a interpretación es sencilla.

Es importante conseguir que los estudiantes que salen de las facultades interpreten correctamente las tablas de contingencia, debido a tres razones: a) su papel en la cultura estadística de todo ciudadano; b) el tema se incluye en las asignaturas de análisis de datos; c) la importancia para la actuación profesional y la investigación en psicología.

El objetivo de este trabajo es hacer un análisis cualitativo exploratorio de las estrategias empleadas en juicios de asociación en estas tablas.

Tabla 1. Ejemplo de tabla

	A	**No A**	**Total**
B	a	b	a+b
No B	c	d	c+d
Total	**a+c**	**b+d**	

INVESTIGACIONES PREVIAS

Los estudios relacionados son numerosos en Psicología, comenzando con Inhelder y Piaget (1955), quienes describen estrategias en el análisis de tablas de contingencia.

El estudio de la precisión en el juicio de asociación ha sido llevado a cabo, entre otros, por Croker (1981), que indica que estos juicios son más exactos si las frecuencias son bajas, se presentan en forma de tablas, los sucesos covarían simultáneamente, las variables son continuas y la correlación es positiva fuerte. Allan y Jenkins (1983) indican que: (1) se tiende a basar los juicios en la diferencia entre casos confirmatorios y no confirmatorios, y (2) la compatibilidad causal entre las variables juega un papel crítico. Erlick y Mills (1967) indican que la asociación negativa se ve próxima a cero. Otros factores que influyen en los juicios de asociación (Arkes y Harkness, 1983) son: (1) la frecuencia en la casilla «*a*» parece tener mayor impacto, (2) la etiquetación de las filas y columnas, y (3) la presencia de números pequeños en las casillas.

Otros autores han estudiado la influencia de las teorías previas en la exactitud de la percepción de la asociación (Jennings, Amabile y Ross, 1982; Wrigth y Murphy, 1984; Alloy y Tabacnick, 1984). La estimación de la asociación es más precisa si las personas no tienen ninguna teoría respecto al tipo de asociación sobre los datos. Si las teorías previas indican el mismo tipo de asociación que hay en los datos empíricos, los sujetos tienden a sobre estimar el coeficiente de asociación. Pero cuando los datos no reflejan los resultados esperados por estas teorías, los sujetos se suelen guiar por sus teorías, más que por los datos. Chapman y Chapman (1969) describen la correlación ilusoria., que consiste en formarse teorías que impide evaluar correctamente las contingencias empíricas. Lleva a la percepción de una relación donde no existe ninguna, o bien a la percepción de una relación más fuerte que la que existente.

La asociación puede ser debida a una relación causa-efecto unilateral, pero también según Barbancho (1973), a la interdependencia, dependencia indirecta, concordancia y covariación espúrea. La comprensión de la asociación implicaría, además de la exactitud en el juicio, comprender estos tipos de relaciones. Sin embargo, Estepa (1993) describe la *concepción causal,* según la cuál el sujeto sólo considera la asociación entre variables si puede adjudicarse a la presencia de una relación causal entre las mismas. También define la *concepción unidireccional* donde el estudiante no admite la asociación inversa, considerándose como independencia.

MÉTODO

La muestra estuvo formada por un total de 62 alumnos de primer año de la Licenciatura en Psicología de la Universidad de Huelva, que cursaban una asignatura de Análisis de Datos I, pero todavía no habían estudiado el tema.

Ítem 1. Se pretende estudiar si el sufrir insomnio tiene relación con los trastornos de estrés. En una muestra de 250 personas observadas se obtuvieron los siguientes resultados:

	Padece estrés	**No padece estrés**
Tener insomnio	90	60
No tener insomnio	60	40

a. Utilizando estos datos razona si ser o no noctámbulo tiene relación con padecer estrés.
b. Indica como has usado los datos de la tabla para llegar a tu conclusión
c. Asigna una puntuación entre 0 (mínimo) y 1 (máximo) según la intensidad de esta relación, marcando una cruz en el punto de esta escala que creas adecuado:

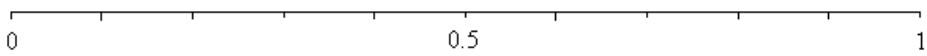


Figura 1. Ejemplo de ítem

El cuestionario que se usó está adaptado del de Estepa y Batanero (1995). En la Figura 1 se presenta el primero de los ítems. El resto son similares, aunque con diferente contexto: ser o no hijo único y ser o no un niño problemático (ítem 2); llevar o no una vida sedentaria y tener o no alergia (ítem 3) y número de horas dedicadas a estudiar un examen y aprobar o suspender (ítem 4, donde el número de horas tiene 3 categoría). Las frecuencias de todas las celdas fueron números menores a 100 y la frecuencia total estuvo comprendida entre 100 y 250 casos. Las variables tenidas en cuenta en el momento de la elaboración del cuestionario (ver tabla 2) fueron las siguientes:

1. *Signo de la asociación entre las variables*, considerándose los tres casos posibles: dependencia directa, dependencia inversa e independencia.
2. *Intensidad de la dependencia,* medida mediante el coeficiente Phi de Pearson en tablas 2x2 y con el coeficiente V de Cramér en tablas 2x3. Se eligieron, un ítem de intensidad moderada-baja y dos de intensidad moderada-alta.
3. *Concordancia entre los datos y las teorías previas sugeridas por el contexto*. Se usaron ítems en que coincide la asociación empírica en los datos con las teorías previas, otros donde no coinciden y un ejemplo no hay teorías previas.
4. *Tipo de covariación*. Utilizamos tres categorías de Barbancho (1973): dependencia causal unilateral, interdependencia, dependencia indirecta.

Tabla 2. Variables de tarea en los ítems

	Tabla 2x2			Tabla 2 x3
	Item 1	**Item 2**	**Item 3**	**Item 4**
Dependencia	Independencia	Inversa	Directa	Directa
Valor del coeficiente de asociación	0	-0,62	0,67	0,37
Concuerda con t. previa	No	Si	No hay teoría	Si
Tipo covariación	Interdependencia	Causal unilateral	Dependencia indirecta	Causal unilateral

RESULTADOS

Dicho estudio semiótico se realiza sobre las respuestas de todos los estudiantes de la muestra a los ítems. Puesto que dichas respuestas son datos textuales, podemos utilizar el análisis de contenido, que se basa en la idea de que las unidades del texto pueden clasificarse en un número reducido de categorías (Krippendorff, 1991). A partir de dicho análisis realizamos inferencias sobre las prácticas matemáticas y conflictos semióticos de los estudiantes mediante la identificación sistemática y objetiva de las características específicas de sus respuestas (Ghiglione y Matalón, 1989).

En lo que sigue analizamos las estrategias, se han considerado las estrategias obtenidas, algunas de las cuáles fueron descritas por Estepa (1993), clasificándolas por correctas, parcialmente correctas e incorrectas.

Estrategias correctas encontradas

ST.1. Comparar todas las distribuciones de frecuencias relativas condicionales de una variable para los distintos valores de la otra variable.

ST.2. Comparación de posibilidades, comparando las frecuencias de casos a favor y en contra de B en cada valor de A.

Estrategias parcialmente correctas

ST.3. Comparar la distribución de frecuencias absolutas condicionales con la frecuencia absoluta marginal correspondiente.

ST.4. Comparar las frecuencias absolutas condicionadas la una con la otra.

ST.5. Comparar la suma de frecuencias en las diagonales.

ST.6. Supone que, para que se la independencia, la frecuencia relativa doble de cada celda de la tabla debe ser igual (es decir el 25% de casos en cada celda).

ST.7. Comparación de posibilidades, comparando las frecuencias de casos a favor y en contra de B en una sola distribución condicional.

Estrategias incorrectas

ST.8. El uso único de la celda de mayor frecuencia.

ST.9. El uso de sólo una distribución condicional.

ST.10. Comparar frecuencias dobles absolutas o relativas entre si o con el número total de observaciones.

ST.11. Comparar frecuencias marginales o medias de las frecuencias marginales.

ST.12. Otros procedimientos incorrectos.

ST.13. El alumno representa gráficamente las distribuciones condicionales por filas, interpretando los datos como si se tratase de observaciones de un análisis de varianza.

ST.14. Se fija en la celda que contradice la asociación.

ST.15. Compara la casilla a con el resto de las celdas, esperando que estas celdas estén en contra de la asociación.

ST.16. El uso de las celdas de mayor y menor frecuencia.

ST.17. El uso de las celdas de menor frecuencia.

ST.18. El uso de tablas de frecuencias marginales.

ST.19. Calculando la diferencia de las diagonales.

ST.20. Comparar los totales de las distribuciones condicionadas.

ST.30. El alumno solo considera sus teorías y no tiene en cuenta los datos.

En la Tabla 3 se resumen las frecuencias de cada una de estas estrategias. La más frecuente resulta ser la estrategia incorrecta, después la estrategia parcialmente correcta y luego la estrategia correcta. Entre las estrategias incorrectas más frecuentes se encuentran el uso de la celda de mayor frecuencia (10), de una sola distribución condicional (11) o comparar frecuencias dobles entre sí (12). Vemos que al ingresar en los estudios de psicología los estudiantes no tienen suficiente capacidad para interpretar una tabla de doble entrada y encontrar la asociación entre las variables. En el caso de Estepa las estrategias más frecuentes fueron: utilizar una única distribución condicional; comparar posibilidades o razón; y comparar una frecuencia relativa de cada distribución.

Tabla 3. Frecuencias de cada estrategia observada

		Ítem 1	Ítem 2	Ítem 3	Ítem 4
Correctas	1	6(9,7)	6(9,7)	6(9,7)	5(8,1)
	2	3(4,8)	1(1,6)		1(1,6)
Parc. correctas	3	4(6,4)		2(3,2)	1(1,6)
	4	8(12,9)	7(11,3)	6(9,7)	19(30,6)
	5		7(11,3)	5(8,1)	2(3,2)
	6*	1(1,6)			
	7*	1(1,6)			
Incorrectas	8	8(12,9)	4(6,5)	6(9,7)	4(6,5)
	9	7(11,3)	11(17,7)	9(14,5)	10(16,1)
	10	8(12,9)	12(19,4)	11(17,7)	4(6,5)
	11	3(4,8)	2(3,2)	2(3,2)	
	12	2(3,2)	3(4,8)	3(4,8)	2(3,2)
	13*	3(4,8)	5(8,1)	5(8,1)	5(8,1)
	14*	1(1,6)		2(3,2)	
	15*	3(4,8)		1(1,6)	
	16*				1(1,6)
	17*		1(1,6)		1(1,6)
	18*				1(1,6)
	19*			1(1,6)	
	20*				1(1,6)
	30*	2(3,2)	1(1,6)	1(1,6)	
	No responde	2(3,2)	2(3,2)	2(3,2)	5(8,1)
Total		**62**	**62**	**62**	**62**

* Estrategias no descritas por Estepa.

CONCLUSIÓN

Aunque la muestra es de tamaño moderado, los participantes fueron un grupo completo de los dos que integran los estudiantes de primer año en la Universidad de Huelva, cuyas características no son muy diferentes de las los estudiantes de psicología en otras universidades españolas, en cuanto a nota de acceso y Bachillerato cursado. No obstante, se deben interpretar los resultados sólo como provisionales y será necesaria una muestra de tamaño mayor.

Al comparar con el estudio de Estepa (1993), se muestran en nuestro caso un menor porcentaje de estrategias correctas para los cuatro ítems (tres casos de tablas 2x2: independencia, asociación inversa, asociación directa; y un caso de tabla 2x3: asociación directa). Sin embargo, la gran mayoría de estrategias son incorrectas (59,3%) mientras que en Estepa, este porcentaje baja al 35,2%, siendo el porcentaje superior al nuestro en estrategias correctas y parcialmente correctas.

En el ítem 1 (independencia), el porcentaje de alumnos que consideran dependencia es aproximadamente el doble que el de alumnos que consideran indepen-

dencia. Pensamos que es debido al mecanismo de la correlación ilusoria, descrito por Chapman y Chapman (1969). Finalmente, habría que comentar que se encontraron estrategias nuevas, respecto a las encontradas por Estepa, dos estrategias parcialmente correctas y nueve incorrectas.

Para finalizar, y puesto que las tablas de contingencia han sido poco estudiadas desde la perspectiva didáctica, consideramos de interés continuar con esta temática.

NOTA DE LOS AUTORES

Agradecimiento a Proyecto EDU2010-1494; Beca FPU-AP2009-2807 (MCINN-FEDER); becas FPU-AP2007-03222 y FPI BES-2008-003573 (MEC-FEDER) y grupo FQM126 (Junta de Andalucía).

REFERENCIAS

Allan, L. G. y Jenkins H. M. (1983). The effect of representations of binary variables on judgment of influence, *Learning and Motivation*, 14, 381-405.

Alloy, L. B. y Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information, *Psychological Review,* 91, 112-149.

Arkes, H. R. y Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology*: General, 112, 1, 117-135.

Barbancho, A. G. (1973). Estadística elemental moderna. Barcelona: Ariel. (Cuarta edición, reimpresión de 1975).

Chapman, L. J. y Chapman, J.P. (1969). Illusory correlation as an obstacle to the use of valid Psychodiagnostic signs, *Journal of Abnormal Psychology,* 74, 271-280.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, 90, 2, 272-292.

Erlick, D.E. y Mills, R.G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology*, 73, 1, 9-14.

Estepa, A. (1993). *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores*. Tesis Doctoral. Universidad de Granada.

Estepa, A. y Batanero, C. (1995). Concepciones iniciales sobre la asociación estadística. *Enseñanza de las Ciencias,* 13(2), 155-170.

Ghiglione, R. y Matalón, B. (1991). *Les enquêtes sociologiques. Théorie et practique.* París: Armand Colin.

Godino, J. D. Batanero, C. y Font, V. (2007). The onto-semiotic approach to research in mathematics education. *ZDM. The International Journal on Mathematics Education, 39* (1-2), 127-135.

Inhelder, B. y Piaget, J. (1955). De la logique de l´enfant à la logique de l´adolescent. París: Presses Universitaires de France.

Jennings, D. L., Amabile, T. M. y Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211-230). Nueva York: Cambridge University Press.

Krippendorff, K. (1991). *Metodología de análisis de contenido. Teoría y práctica.* Barcelona: Paidos.

Wright, J. C. y Murphy, G.L. (1984). The utility of theories in intuitive statistics: the robustness of theory-based judgments, *Journal of Experimental Psychology General,* 113(2), 301-322.

APLICACIÓN DE LA REGRESIÓN LOGÍSTICA A LA DETECCIÓN DE FACTORES DE INFLUENCIA EN EL NIVEL DE ACTIVIDAD FÍSICA DE LOS ADOLESCENTES

Carlos A. Cordente-Martínez[1] y Pilar García-Soidán[2]

[1] Universidad Politécnica de Madrid

[2] Universidad de Vigo

Correo electrónico: cordente@upm.es

Resumen

Diversos estudios de investigación, llevados a cabo en las dos últimas décadas, aportan datos preocupantes sobre la escasa actividad física que desarrollan los niños/as y adolescentes. Por esta razón, en el presente trabajo se trataron de identificar aquellos factores (sociales, biológicos y psicológicos) que han dado lugar a esta realidad y que, por tanto, tienen mayor peso en el nivel de actividad física que realizan las personas jóvenes. Para ello se han utilizado los datos recogidos en diversos centros de educación secundaria (públicos y privados) de los distintos distritos del municipio de Madrid y se han aplicado las técnicas de regresión logística, obteniéndose un modelo de predicción del nivel de actividad física a partir de los aspectos más influyentes, cuyos resultados más significativos se resumen en el presente trabajo.

En los últimos años se ha observado en España un incremento preocupante del número de personas que padecen enfermedades relacionadas con el excesivo sedentarismo, entre las que se incluyen las afecciones cardiovasculares, la obesidad, etc. Esta realidad ha supuesto que se dedique mayor atención a la medicina preventiva y que se promocione la realización de actividad física de forma regular y de un estilo de vida más saludable. Diversos estudios de investigación (Cantera y Devís, 2002; Janz et al., 2000; Sallis et al., 2000) muestran datos preocupantes sobre la escasa actividad física que desarrollan los niños/as y adolescentes e interesa analizar las causas que han dado lugar a esta realidad.

Teniendo presente lo anterior, en este trabajo se han tratado de identificar aquellos factores (sociales, biológicos y psicológicos) que tienen mayor peso en el nivel de actividad física que realizan los adolescentes españoles. Para ello se han utilizado los datos recogidos en diversos centros de educación secundaria (públicos y privados) de los distintos distritos del municipio de Madrid y se han aplicado las técnicas de regresión logística, obteniéndose un modelo de predicción del nivel de actividad física a partir de los aspectos más influyentes.

Entre los resultados obtenidos se observa que los factores sociales tienen un peso importante en la actitud de las personas jóvenes hacia la práctica deportiva, destacando principalmente la realización de actividad física en su entorno más próximo, particularmente sus amistades y progenitores. También intervienen factores biológicos, como el sobrepeso, que preocupa particularmente a las autoridades sanitarias por incrementar el riesgo de padecer enfermedades coronarias, diabetes, etc. Además, la población adulta con problemas de obesidad está aumentando de forma alarmante en los países desarrollados y entre las causas que han dado lugar a esta situación se destacan, entre otras, algunas ligadas a malos hábitos o problemas heredados de la etapa adolescente, como el sobrepeso, la sobrealimentación y la inactividad física.

MÉTODO

Diseño del estudio

Para desarrollar este estudio se han utilizado los datos de 554 estudiantes de diversos centros de educación secundaria (18 públicos y 17 privados) de 16 distritos del municipio de Madrid. Con este tamaño muestral, el error máximo admisible es de ±5,6%, al nivel de confianza del 95%.

Para medir el nivel de actividad física de los participantes en el estudio se utilizó el Modifiable Questionnaire of Physical Activity for adolescents (Aaron y Kriska, 1997), que previamente había sido traducido y validado para su uso en España (Cordente, 2006). Para la clasificación del nivel de actividad física se tuvieron presentes las recomendaciones de las autoridades inglesas en materia de educación para la salud (Pate et al., 1998) del modo siguiente:

- >49 METs/Semana = Adolescente activo/muy activo.
- 8-49 METs/ Semana = Adolescente moderadamente activo.
- 28 METs/Semana = Adolescente inactivo/sedentario.

Los datos recogidos fueron analizados con el programa estadístico SPSS. De este modo se realizó en primer lugar un análisis exploratorio para determinar las asociaciones existentes entre las diferentes variables consideradas. A continuación se aplicó el método de regresión logística, con objeto de establecer un modelo que permitiese predecir el nivel de actividad física de un adolescente (NAF) a partir de los valores observados en los factores asociados con el mismo. La codificación de las variables utilizadas se resume en la tabla 1.

Tabla 1. Variables consideradas en el estudio

Variable	Codificación
Nivel de actividad física (NAF)	0 (inactivo/moderadamente activo) y 1 (activo/muy activo)
Género	0 (mujer) y 1 (hombre)
Tipo de centro educativo	0 (público) y 1 (privado)
El padre realiza actividad física (PAF)	0 (no) y 1 (sí)
La madre realiza actividad física (MAF)	0 (no) y 1 (sí)
Los amigos realizan actividad física (AAF)	0 (no) y 1 (sí)
Sobrepeso	0 (sí) y 1 (no)
Consumo de tabaco	0 (no) y 1 (sí)
Consumo de alcohol	0 (no) y 1 (sí)

PRINCIPALES RESULTADOS Y DISCUSIÓN

El nivel de actividad física de los adolescentes participantes en el estudio se resume en el gráfico siguiente.

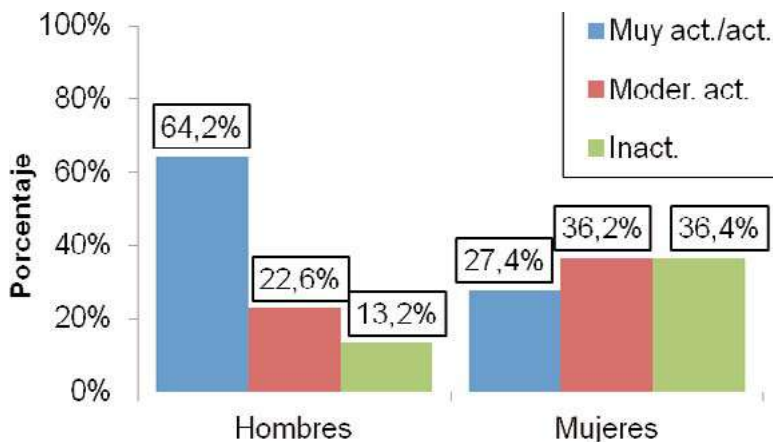


Figura 1. Distribución de la muestra por NAF y género

Para establecer el grado de asociación entre el NAF de los adolescentes y las restantes variables consideradas, se aplicó el test chi-cuadrado, cuyos resultados se presentan a continuación.

Tabla 2. Grado de asociación entre el NAF y las restantes variables

Variable	Test chi-cuadrado	Significación
Género	75,852	0,000
Tipo de centro educativo	7,026	0,008
PAF	10,555	0,001
MAF	3,297	0,069
AAF	14,805	0,000
Sobrepeso	23,236	0,000
Consumo de tabaco	1,419	0,234
Consumo de alcohol	0,528	0,468

De la tabla anterior se deduce que, al nivel de confianza del 95%, existe una importante asociación entre el NAF y las siguientes variables (significación < 0,05): género, sobrepeso, PAF y AAF.

Seguidamente se aplicó el análisis de regresión logística, tomando como variable dependiente NAF y como independientes las restantes. Para la introducción de variables se utilizó el criterio de entrada de Wald, finalizando el proceso cuando ninguna variable podía ser introducida o eliminada de modo que mejorase significativamente el ajuste. Los 4 factores incorporados al modelo resultante fueron precisamente los que el test chi-cuadrado revelaba como los más influyentes. Los coeficientes estimados obtenidos en este ajuste y su grado de significación se presentan en la tabla 3, de los cuales se deduce que ningún coeficiente es nulo (significación < 0,05) y, por tanto, no se puede prescindir de las variables introducidas.

Tabla 3. Coeficientes del modelo de regresión logística (estimación y significación)

Variable	Parámetro	Test de Wald	Significación
Género	1,405	53,886	0,000
PAF	0,536	7,529	0,006
MAF	0,864	8,893	0,003
AAF	1,153	7,024	0,008
Sobrepeso	-2,912	34,746	0,000
Constante	1,405	53,886	0,000

En consecuencia, la regresión logística nos conduce al modelo siguiente:

Tabla 4. Modelo de regresión logística resultante

P0 = Probabilidad de que un adolescente sea moder. activo/inactivo = 1 / (1+eZ)
P1 = Probabilidad de que un adolescente sea muy activo/activo = eZ / (1+eZ)
donde:

$$Z = -2.9 + 1.405 \times \text{Género} + 0.536 \times \text{PAF} + 0.864 \times \text{AAF} + 1.153 \times \text{Sobrepeso}$$

Con las 4 variables incluidas en el modelo, la matriz de confusión, que proporciona los resultados de la clasificación, tomaría los valores indicados en la tabla 4, de los cuales se deduce que el modelo resultante tiene una probabilidad de acierto para casi el 70% de los casos.

Tabla 5. Clasificación resultante del modelo de regresión logística

NAF observado	NAF pronosticado		% correcto
	Moder. act./inact.	Muy act./act.	
Moder. act./inact.	88	81	**73,4%**
Muy act./act.	0,536	162	**64,8%**
% global			69,5%

CONCLUSIONES

La práctica de la actividad física está condicionada por múltiples factores. Desde esta perspectiva, nuestra principal aportación ha sido tratar de detectar las variables más influyentes en la población adolescente. Para ello, hemos ajustado un modelo que permite predecir el NAF a partir de los factores más relacionados y que han resultado los siguientes (en orden decreciente de importancia): género, AAF, PAF y sobrepeso.

Con excepción del género, los demás factores son modificables, de ahí que consideramos que se deberían tener presentes a la hora de diseñar campañas de promoción de la actividad física entre los adolescentes, particularmente para las mujeres, con una práctica de actividad física claramente inferior y a las que se debería prestar atención prioritaria para la promoción de un estilo de vida activo. La fiabilidad del ajuste resultante (casi un 70%) podría optimizarse en estudios futuros incluyendo otras variables que pueden condicionar la práctica de actividad física, como la disponibilidad de infraestructuras deportivas o el nivel de inseguridad en la zona de residencia.

REFERENCIAS

Aaron, D.J. & Kriska, A.M. (1997). Modifiable activity questionnaire for adolescents. *Medicine & Science in Sports & Exercise, 29,* 79-82.

Cantera, M.A. & Devís-Devís, J. (2002) La promoción de la actividad física relacionada con la salud en el ámbito escolar. Implicaciones y propuestas a partir de un estudio realizado entre adolescentes. *Apunts. Educación Física y Deportes, 67*, 54-62.

Cordente Martínez, C.A. (2006) *Estudio epidemiológico del nivel de actividad física y de otros parámetros de interés relacionados con la salud bio-psico-social de los alumnos de ESO del Municipio de Madrid*. Toledo: Universidad de Castilla-La Mancha.

Janz, K.F., Dawson, J.D. & Mahoney, L.T. (2000). Tracking physical fitness and physical activity from childhood to adolescence: the Muscatine Study. *Medicine & Science in Sports & Exercise, 32*, 1250-1257.

Pate, R.R., Trost. S. & Williams, C. (1998). Critique of existing guidelines for physical activity in young people. In: Biddle, S., Sallis, J. & Cavill, N. (ed.) *Young and active? Young people and health enhancing physical activity: evidence and implication.* London: Health Education Authority, 162-176.

Sallis, J.F., Prochaska, J.J. & Taylor, W.C. (2000). A review of correlates of physical activity of children and adolescents. *Medicine & Science in Sports & Exercise, 32,* 963-975.

SIMPLIFICAR LA CONVERSIÓN DE DATOS OBSERVACIONALES EN EL DEPORTE CON EL SOFTWARE LINCE

Brais Gabin[1], Oleguer Camerino[2] y M. Teresa Anguera[3]
[1] Universidad Politécnica de Barcelona
[2] Universitat de Lleida
[3] Universidad de Barcelona
Correo electrónico: ocamerino@inefc.es

Resumen

Actualmente para aplicar procedimientos de observación informatizados en el deporte nos hemos enfrentado a numerosos problemas derivados de la inexistente versatilidad de los datos, las dificultades en el registro de imágenes, y la falta de inmediatez en los resultados. El software Lince ha sido desarrollado por un equipo de ingenieros informáticos y especialistas en observación -Proyecto de I+D+I del Ministerio de Ciencia e Innovación: A*vances tecnológicos y metodológicos en la automatización de estudios observacionales en deporte* (Grant PSI2008-01179)- con el objetivo de superar los problemas del registro, cómputo y conversión de datos y facilitar el trabajo de los investigadores del deporte. La aplicación se ha desarrollado en Java y tiene una interfaz ágil, intuitiva, útil y veloz para que sea una aplicación multiplataforma para que pueda ser ejecutada en diferentes sistemas operativos. Además esta aplicación se desarrollara bajo una licencia GNU GPLv3 (GNU General Public License) que garantizara la libertad del código. Esto permite su libre distribución (http://lom.observesport.com/) y la mejora por parte de cualquier persona que así lo desee permitiendo esto que la aplicación vaya ampliando en funcionalidades.

La aplicación LINCE ha sido diseñada para facilitar la tarea observacional y puede complementarse a otros software de similares características (Castellano, 2005, 2008; González, Hernández Mendo y Pastrana, 2010). La aplicación trata de superar todas las debilidades de los programas actuales de observación: necesidad de acciones repetitivas, formatos propios que requieren ser transformados para volcarlos en otros programas, difícil manejo de las acciones de la observación, limitaciones en la capacidad del número de criterios o categorías, imposibilidad de poder modificar o añadir nuevos módulos para fines específicos, y limitaciones en los formatos de vídeo a visionar.

Para poder conseguir unos objetivos tan ambiciosos la primera característica de LINCE es que es software libre. Es gratuito y se distribuye tanto la aplicación como

su código. Con esto se consigue que cualquier persona con ciertos conocimientos en programación pueda modificar, mejorar o añadir cualquier funcionalidad para adecuarlo a sus propias necesidades y que su trabajo beneficie a los demás usuarios. LINCE ha generado toda su documentación bajo una Licencia Creative Commons BY-NC-SA para facilitar tanto el uso como la modificación de LINCE.

La aplicación está desarrollada en el lenguaje de programación Java que permite tener una aplicación multiplataforma en sistemas operativos actuales (i.e. Windows, Linux, Macintosh) esta característica es cada vez más requerida por los usuarios dado que en los últimos años la diversidad en el uso de sistemas operativos ha aumentado vertiginosamente.

LINCE (ver fig. 1) permite observar cualquier episodio de actividad física o deporte al estar construido como un paquete informático para automatizar las funciones de: diseño de sistemas observacionales, registro en vídeo, cálculo de la calidad del dato y presentación de resultados para su exportación en diversos formatos: TxT, Theme, GSEQ, Excel y SAS.

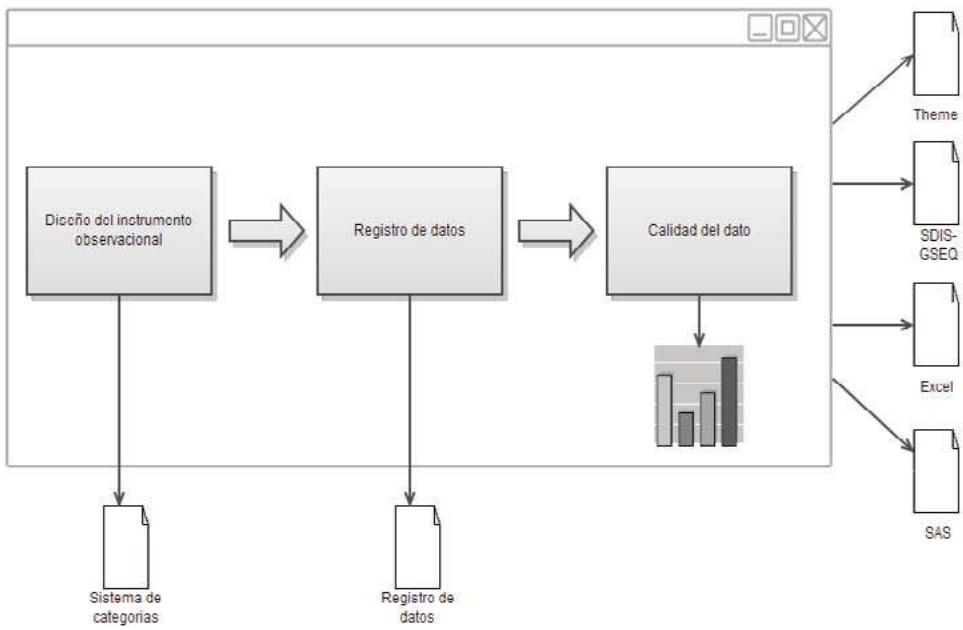


Figura 1. Gráfico de las funciones de LINCE.

FUNCIONALIDADES DE LINCE

Este instrumento supera algunos de los problemas clásicos en la obtención de registros y el cómputo de datos observacionales del ámbito de las Ciencias de la Actividad Física y el Deporte:

Instalación en una sola operación de descarga. El proceso instantáneo aporta todos los requerimientos necesarios para su funcionamiento: maquina virtual Java y reproductor VLC que puede trabajar con la mayoría de archivos multimedia.

Construcción de instrumentos de observación. Permite un número ilimitado de criterios fijos, mixtos y variables. Estos últimos permiten incluir tantos niveles de categorías y subcategorías como se desee. Todos los tipos de criterios pueden ser modificados sin alterar los registros realizados hasta el momento. Esto permite cambios y ajustes en los momentos iniciales de la determinación del instrumento observacional (Anguera, 2003).

Visualización de imágenes. Podemos cargar cualquier tipo de imágenes (incluyendo actuales formatos de HD) y reproducirlas con una precisión de milésimas de segundo. El control de esta reproducción se puede hacer mediante diversas opciones para adaptarse mejor a la forma de trabajo de cada usuario. Es posible controlar con una botonera, atajos de teclado o incluso con la pulsación de los botones centrales y derecho del ratón (ver figura 2).

Control del registro. El registro de acontecimientos, que puede ser modificado, queda constatado simultáneamente con el tiempo y su duración en segundos o *frames* permitiendo así estudios diacrónicos con descripción de secuencias con multiacontecimiento.

Figura 2. Interface del programa con ejemplo de observación de judo

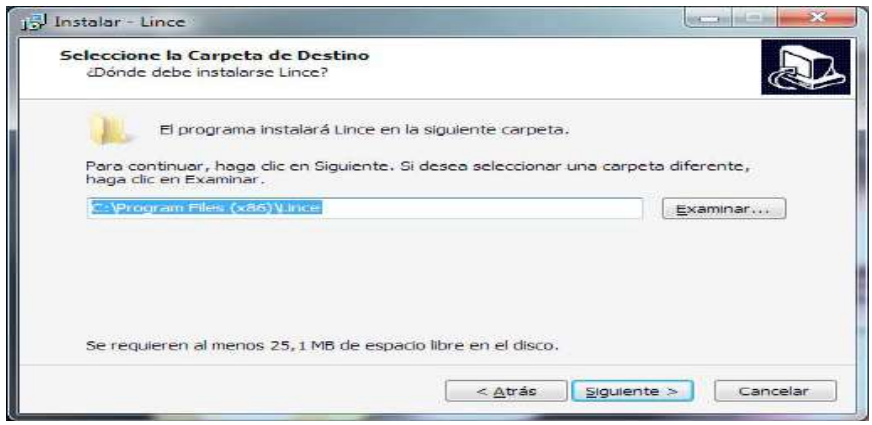


Figura 3. Pantalla del instalador de LINCE

Cálculo de la calidad del dato. LINCE permite hacer cálculos del coeficiente kappa de Cohen (Cohen, 1960), de todos los criterios o de algunos de ellos, mediante la comparación de dos archivos de registro del mismo observador (fiabilidad intraobservador) y de varios observadores diferentes (fiabilidad interobservador).

Versatilidad en la exportación de los datos. Uno de los puntos clave de LINCE es la posibilidad de exportar los datos a otros formatos para su posterior tratamiento. LINCE exporta a formatos específicos de los programas más usados en el cálculo de los resultados observacionales y también a formatos genéricos como puede ser .csv el cual es posible abrir con Excel del paquete ofimático Office de Microsoft o Calc de OpenOffice.

INSTALACIÓN DE LINCE

Tras descargar el instalador de Lince (http://lom.observesport.com) podemos pasar a su instalación. Para esto primero ejecutamos el .exe descargado, vamos pulsando el botón de siguiente aceptando la licencia y decidiendo el lugar donde queremos instalar el programa (recomendado no modificar la ruta que viene por defecto) y si queremos un acceso directo en nuestro escritorio. Tras estos pasos se instalará Lince. Junto con Lince se nos actualizará la maquina virtual de Java y el reproductor multimedia VLC, ambos requisitos para que Lince funcione correctamente. (ver Fig 3).

Desde el menú inicio de Windows se puede acceder tanto al programa LINCE como al manual de usuario. En el manual de usuario se explica cómo utilizar LINCE paso por paso. Que muestra cada una de las pantallas, como utilizarlas y que se puede hacer desde cada una de ellas.

EXPORTACIÓN DE LOS DATOS DE LINCE

Los datos que se obtienen mediante LINCE pueden ser posteriormente exportados a diferentes aplicaciones de análisis. Esto se consigue gracias a que LINCE reconoce los formatos de los archivos de las principales aplicaciones de análisis: análisis de secuencia de retardos GSEQ (Bakeman & Quera, 1996), cálculo de la varianza SAS (Ysewijn, 1996; Schlotzhauer & Littell, 1997), análisis secuencial de T-patterns Theme (Magnusson, 1996, 2000, 2005) y indicadores estadísticos descriptivos y correlacionales con Excel.

Actualmente LINCE puede exportar a ocho formatos diferentes –5 formatos de Sdis, theme, Txt y Excel– que irá creciendo en las sucesivas versiones (fig 4, 5 y 6).

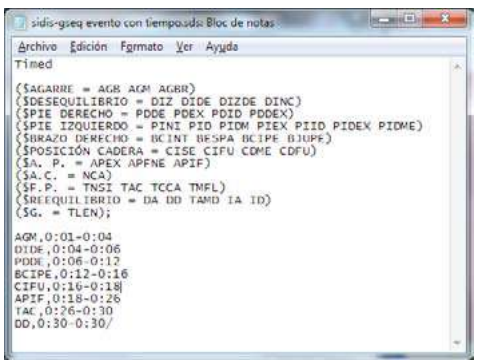


Figura 4. Pantalla del formato Sgsq

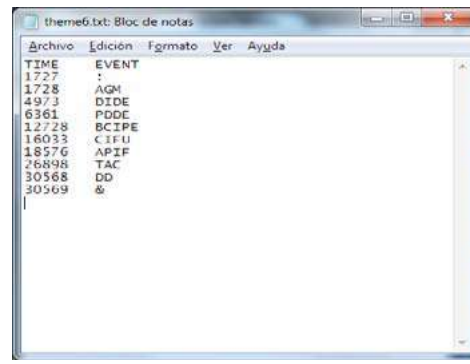


Figura 5. Pantalla del formato THEME

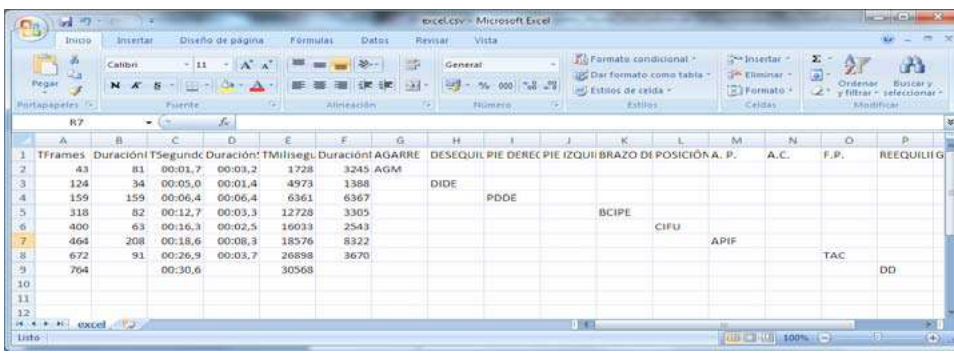


Figura 6. Pantalla del formato EXCEL

Además los datos a exportar son completamente configurables, lo que significa que podemos especificar de una forma muy sencilla que criterios queremos añadir y cuales no.

CONCLUSIÓN

Lince es una aplicación gratuita y fácil de manejar que permite seguir todos los pasos de la metodología observacional con suma eficiencia. Sus funcionalidades están hechas para ahorrar trabajo de los investigadores de la observación de la actividad física y el deporte implementando la calidad del dato y la exportación de una forma nativa.

Consideramos que Lince es una aplicación que no impone límites: no hay número máximo de criterios, no hay restricciones en los formatos de video y no hay límite en el tamaño del registro.

Además Lince es un programa que acaba de nacer, sigue en desarrollo para aumentar sus funcionalidades por lo que todo feedback es bienvenido como podrían ser: peticiones, consejos, reportes de fallos y sugerencias.

LINCE software puede ayudar a la obtención de datos observacionales de los episodios deportivos que ayuden a la mejora de esta mayor comprensión de los fenómenos complejos y dinámicos que son responsables de esta eficiencia competitiva.

NOTA DE LOS AUTORES

Agradecemos el soporte del Gobierno Español en el proyecto que subvenciona *Innovaciones en la evaluación de contextos naturales: Aplicaciones al ámbito del deporte* (Dirección General de Investigación, Ministerio de Ciencia y Tecnología) [(PSI 2008-01179)].

REFERENCES

Anguera, M.T. (2003). Observational Methods (General). In R. Fernández-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment, Vol. 2* (pp. 632-637). London: Sage.

Bakeman, R., & Quera, V. (1992). SDIS: A sequential data interchange standard. *Behavior Research Methods, Instruments & Computers, 24*, 554-559.

Castellano, J., Perea, A., & Alday, L. (2005). Match Vision Studio v3.0. Paper presented at Measuring Behavior 2005. 5th International Conference on Methods and Techniques in Behavioral Research. Wageningen, The Netherlands.

Castellano, J., Perea, A., Alday, L., & Hernandez-Mendo, A. (2008). The Measuring and Observation Tool in Sports. *Behavior Research Methods* 2008, 40 (3), 898-905

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin, 70*, 213-220.

González Ruiz, S. L., Hernández Mendo, A., Pastrana Brincones, J.L. (2010). Herramienta software para la evaluación psicosocial de deportistas y entornos deportivos. Lecturas: EF y Deportes. Revista Digital,15(144), mayo.

Magnusson, M.S. (1996). Hidden real-time patterns in intra- and inter-individual behavior. *European Journal of Psychological Assessment, 12* (2), 112-123.

Magnusson, M.S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers, 32* (1), 93-110.

Magnusson, M.S. (2005) Understanding Social Interaction: Discovering Hidden Structure with Model and Algorithms. In L. Anolli, S. Duncan, M. Magnusson & G. Riva (Eds.), *The hidden structure of social interaction. From Genomics to Culture Patterns*. (pp 51-70). Amsterdam: IOS Press.

Schlotzhauer, S. D., & Littell, R. C. (1997). SAS system for elementary statistical analysis (2nd ed.). Cary, NC: SAS Institute.

Ysewijn, P. (1996). GT: Software for generalizability studies. Available online at www.irdp.ch/methodo/generali.htm.

ESTUDIO INMA (INFANCIA Y MEDIO AMBIENTE): DISEÑO LONGITUDINAL DE COHORTE

Nerea Lertxundi [1,4], Eduardo Fano[1,4], Aitana Lertxundi[1,3,4], Oscar Vegas[1,4], Aritz Aranbarri[1,4], Ainara Andiarena[1] y Jesus Ibarluzea[2,3,4]

[1] Universidad del País Vasco

[2] Subdirección de Salud Pública de Gipuzkoa

[3] Consorcio de Investigación Biomédica de Epidemiología y Salud Pública (CIBERESP)

[4] Instituto de Investigación Sanitaria BioDonostia

Correo electrónico: nerea.lertxundi@ehu.es

Abstract

Se presenta el diseño de investigación del estudio INMA (Infancia y Medio Ambiente), cuyo objetivo general es estudiar el papel de los contaminantes ambientales más importantes en el aire, agua y en la dieta durante el embarazo e inicio de la vida y sus efectos en el crecimiento y desarrollo neuropsicológico infantil. La Red INMA se constituyó en 2003 y está formada por siete cohortes que están llevando a cabo el seguimiento de aproximadamente 4.000 mujeres embarazadas y sus hijos/as. Se trata de un diseño longitudinal de cohorte. La recogida de datos se ha realizado en los tres trimestres del embarazo, en el nacimiento, a los 14 meses de edad de los niños/as, a los 2 años y 4 años y 4 meses. El proyecto permite disponer de información a nivel individual (historia clínica, cuestionarios, pruebas físicas y neuropsicológicas y biomarcadores de exposición o efecto) e información a nivel ecológico (nivel de contaminantes en el medio ambiente, características urbanísticas) a lo largo del tiempo, de forma que es posible relacionar exposiciones ambientales con posibles efectos en el desarrollo físico y neuropsicológico.

El desarrollo físico, social e intelectual de los niños/as desde la concepción hasta el final de la adolescencia requiere de un ambiente protegido. Un número creciente de enfermedades en los niños/as están vinculados a las condiciones de inseguridad. Las exposiciones de la vida prenatal y temprana están asociadas a la predisposición de efectos sobre la salud durante el desarrollo infantil y la vida adulta. El proyecto INMA –Infancia y Medio Ambiente– (Infancia y Medio Ambiente) es una red de proyectos de cohortes de nacimiento en España que tienen como objetivo estudiar el papel de los contaminantes ambientales en el aire, agua y la dieta durante el embarazo y la primera infancia en relación con el crecimiento y desarrollo infantil. El proyecto, en su conjunto, surgió como parte de las Redes Temáticas de Investigación Cooperativa (Red INMA: acrónimo de Infancia y Medio Ambiente), creadas por el Ministerio de Sanidad y Consumo (Instituto de Salud

Carlos III), en buena medida, para dar respuesta a la necesidad plantada por la Comunidad Europea de estudiar los factores ambientales y su impacto en la infancia. Actualmente, la gran mayoría de los equipos de investigación, forman parte del CIBER de Epidemiología y Salud Pública.

INMA se basó en las experiencias adquiridas por tres cohortes de nacimiento de la cohorte de Ribera d'Ebre (n = 102), que evaluó el desarrollo neurológico en un área con altos niveles de compuestos organoclorados (OC) y de mercurio procedentes de la emisión de una planta electroquímica, la cohorte de Menorca (n = 530), que estudió la asociación entre la exposición temprana a contaminantes atmosféricos y efectos sobre las vías respiratorias, alergias y el asma, y la cohorte de Granada (n = 668), que estudiaron la incidencia de los trastornos del bebé al nacer, la salud reproductiva en relación a los disruptores endocrinos ambientales. Basándose en la experiencia de estas cohortes, se desarrolló un nuevo protocolo común para cuatro nuevas cohortes: Valencia (n = 855), Sabadell (n = 657), Asturias (n = 494) y Gipuzkoa (n = 638) (Guxens, 2011).

Los objetivos generales del estudio son:

1. Describir el grado de exposición individual prenatal a contaminantes ambientales y la dosis interna de productos químicos durante el embarazo, el parto y durante la infancia en España.
2. Evaluar el impacto de la exposición a diferentes contaminantes en el crecimiento fetal e infantil, la salud y el desarrollo.
3. Evaluar la interacción entre los contaminantes persistentes, los nutrientes, y las variables genéticas en el crecimiento fetal e infantil, salud y desarrollo.

Las características específicas de cada una de las cohortes así como las diferentes etapas de seguimiento explican los objetivos específicos que se trabajan en cada una de las etapas del estudio. A modo de ejemplo, se detallan los objetivos específicos de la cohorte de Gipuzkoa en la fase de los 4 años:

1. Evaluar el papel de la exposición a contaminación atmosférica por partículas en el periodo prenatal y primeros años de infancia en la incidencia de asma y en la función pulmonar a los 4 años de edad.
2. Evaluar el papel de la exposición a contaminación atmosférica en el periodo prenatal y primeros años de vida en el desarrollo mental a los 4 años de edad.
3. Evaluar el papel de los niveles de plaguicidas (DDT, DDE, HCH, HCB y PCBs) en sangre de madre durante el embarazo y cordón umbilical en el desarrollo mental a los 4 años.
4. Evaluar el papel modificador de los genes metabolizadores.
5. Estudiar el papel protector de los antioxidantes de la dieta, incorporados en la vida en el periodo prenatal y actual, en las asociaciones anteriores.

6. Estudiar el papel modificador de la calidad del contexto familiar, coeficiente intelectual y vínculo afectivo de los padres en el desarrollo mental a los 4 años.
7. Estudiar el papel modificador de los indicadores de funcionamiento del sistema hipotálamo-hipófisis-adrenal (cortisol) y simpático-adreno-medular (alfa-amilasa).

DISEÑO Y MÉTODO

Se trata de un diseño longitudinal de cohorte. Los criterios para la inclusión de las madres fueron: (i) que fueran residentes en una de las áreas de estudio, (ii) mínimo de edad de 16 años, (iii) tener un embarazo único, (iv) no haber seguido un programa de reproducción asistida, (v) dar a luz en el hospital de referencia y (vi) no tener problemas de comunicación. Cada cohorte comenzó con la primera fase de reclutamiento en diferentes momentos (Figura 1). En Ribera d'Ebre y las cohortes de Granada, las madres fueron reclutadas durante el ingreso hospitalario. El reclutamiento se llevó a cabo entre marzo de 1997 y diciembre de 1999 (Ribera d'Ebre) y octubre de 2000 y julio de 2002 en la cohorte de Granada. En la cohorte de Menorca, todas las mujeres embarazadas que se presentaron para la atención prenatal en toda la isla (en centros de salud pública o privada) fueron invitadas a participar en el estudio sobre un período de 12 meses a partir de mediados de 1997. En las cuatro nuevas cohortes, el reclutamiento se llevó a cabo durante la primera visita prenatal (10-13 semanas de gestación) en los principales hospitales públicos o centros de salud de cada área de estudio. El período de reclutamiento en Valencia fue de noviembre de 2003 a junio de 2005. En Sabadell, entre julio de 2004 y julio de 2006. El reclutamiento en en Asturias se llevó a cabo entre mayo de 2004 y julio de 2007, y en Gipuzkoa desde abril de 2006 a enero de 2008.

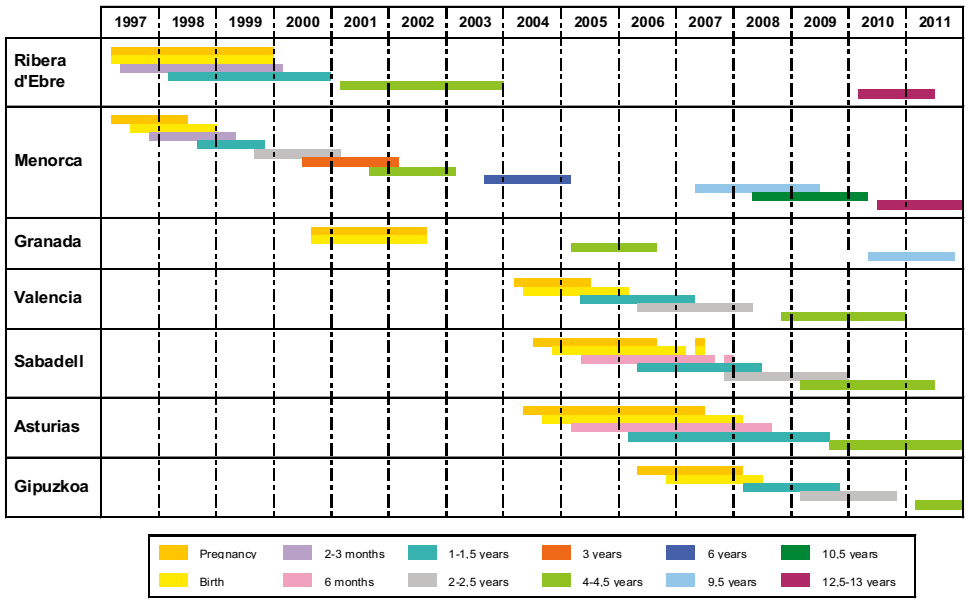


Figura 1. Etapas de seguimiento de las cohortes INMA

Una característica común de estas áreas es que la gran mayoría de la población asiste al servicio público de salud. Respecto a los datos de los niveles de participación, cabe señalar que entre el 45% y 98% de las mujeres embarazadas aceptaron participar en el estudio (96% en Ribera d'Ebre, el 98% en Menorca, el 54% en Valencia, el 60% en Sabadell, el 45% en Asturias y el 68% en Gipuzkoa, información no disponible para Granada). En Sabadell el nivel educativo de las mujeres que decidieron no participar en el estudio fue menor que el de las mujeres participantes, mientras que no se encontraron diferencias respecto a la edad. En Valencia, una proporción de mujeres mayores y de mujeres que trabajaban aceptaron participar en el estudio. No hubo diferencias de edad entre participantes y no participantes en Asturias. En Gipuzkoa, una alta proporción de mujeres que trabajaban fueron incluidas. La información comparativa de las características básicas de las participantes y no participantes no está disponible para Ribera d'Ebre, Menorca y Granada. No obstante, dos cohortes (Sabadell y Gipuzkoa) recopilaron información sobre las razones por las que decidieron no participar en el estudio: 27,6% lo hizo porque no deseaba participar, el 30,3% dijo que no tenía tiempo, el 9,3% por falta de interés y el 32,7% porque no fueron localizadas para la primera entrevista (Guxens, 2011).

Tabla 1. Información recogida en cada una de las etapas del seguimiento.

	Periodo Prenatal			Periodo Postnatal		
	12 semanas	20 semanas	32 semanas	Nacimiento	1 año	4 años
Exposiciones						
PAHs, VOCs. PMs y NO2	Exterior		Interior (submuestra) Cuestionario		Cuestionario	Cuestionario
Trihalometanos	Exterior		Interior (submuestra) Cuestionario		Cuestionario	Cuestionario
Ocupación madre			Cuestionario			
OCs/Policromados PBDDs y ftalatos	Suero materno			Suero cordón umbilical		Suero niño
Disruptores endocrinos				Placenta		
Plomo				Sangre cordón umbilical		Suero niño
Arsénico	Uña madre					Uña niño
Metil-mercurio				Pelo recién nacido		Pelo niño
Hidroxipireno	Orina madre					Orina niño
Dieta	Cuestionario		Cuestionario		Cuestionario	Cuestionario
Antioxid. Liposolubles Ac. Ascórbico CAT Folatos	Suero materno					
Acidos Grasos	Suero materno			Leche materna		
Efectos						
Estrés oxidativo	Suero materno		Orina madre			Suero niño Orina niño
Crecimiento intrauterino	Ecografía	Ecografía	Ecografía			
Desarrollo sexual				Exploración física	Exploración física	Exploración física
Crecimiento postnatal				Exploración física	Exploración física	Exploración física
Neurodesarrollo				Test de Dubowitz	Test de Bayley	Test de McCarthy Hiperactividad Competencia Social
Inmunidad	Suero materno				Cuestionario	Suero niño
Hormonas Tiroideas	Suero materno			TSH screening		Suero de niño
Otros						
Estudio genético	Sangre materna			Sangre de cordón		
Psicoafectividad padres					Cuestionario de apego Cuestionario de salud mental	

La información se recoge según el protocolo establecido por el Proyecto INMA (www.proyectoinma.org). La recogida de datos se ha realizado en los tres trimestres del embarazo, en el nacimiento, a los 14 meses de edad de los niños/as, a los 2 años (solo en Gipuzkoa) y a los 4 años y 4 meses (Tabla 1). Se ha obtenido información sobre datos sociodemográficos, de residencia, hábitos de vida, ocio, historia clínica, historia de reproducción e información sobre la dieta través de cuestionarios estandarizados administrados por personal entrenado del INMA. También se han realizado evaluaciones físicas, psicológicas, neuropsicológicas y del entorno familiar y se han recogido muestras biológicas (sangre, placenta,cordón umbilical, calostro, pelo, uña, orina, saliva) para la cuantificación de biomarcadores de exposición y parámetros clínicos o bioquímicos. Para la obtención de datos sobre los diferentes contaminantes ambientales se han utilizado programas establecidos por la administración o procedimientos de muestreo «ad hoc» con el fin de recabar información sobre los niveles de contaminantes de interés en el aire exterior e interior, en el agua de consumo y en el baño de las áreas de estudio de la muestra.

RESULTADOS

El proyecto permite disponer de información a nivel individual (historia clínica, cuestionarios, pruebas físicas y neuropsicológicas y biomarcadores de exposición o efecto) e información a nivel ecológico (nivel de contaminantes en el medio ambiente, características urbanísticas) a lo largo del tiempo, de forma que se relacionan exposiciones ambientales con posibles efectos en el desarrollo físico y neuropsicológico. El estudio INMA sigue recogiendo información y realizando los análisis de datos disponibles hasta la actualidad en las 4 cohortes nuevas. Las referencias de las publicaciones están en la página web con enlace a Pubmed (http://www.proyectoinma.org/presentacion-inma/resultats_en). Algunos de los resultados publicados concluyen que la exposición a varios de los contaminantes del medio ambiente se relacionan con la alteración de los niveles de las hormonas tiroideas, tanto maternas como infantiles (Álvarez-Pedrerol, 2009; López-Espinosa, 2010; López-Espinosa, 2009), con un mayor riesgo de contraer anomalías congénitas (Ribas-Fitò, 2002; Fernández, 2007), con el deterioro del feto y problemas de crecimiento de los niños/as (Aguilera, 2009; Aguilera, 2010; Ballester, 2010; Mendez, 2008; mendez, 2011; Llop, 2010), con el retraso en el desarrollo cognitivo y comportamental del niño/a (Freire, 2010, Julvez, 2007; Vrijheid, 2010) y con un mayor riesgo de problemas respiratorios (Friguls, 2009; Polk, 2004; Torrent, 2006; Torrent, 2007; Sunyer, 2010).

REFERENCIAS

Aguilera, I., Guxens, M., Garcia-Esteban, R. et al. (2009). Association between GIS-based exposure to urban air pollution during pregnancy and birth weight in the INMA Sabadell Cohort. *Environ Health Perspect*,117, 1322–27.

Aguilera, I., Garcia-Esteban, R., Iniguez, C. et al. (2010). Prenatal exposure to traffic-related air pollution and ultrasound measures of fetal growth in the INMA-Sabadell cohort. *Environ Health Perspect,* 118, 705–11.

Alvarez-Pedrerol, M., Guxens, M., Ibarluzea, J. et al. (2009). Organochlorine compounds, iodine intake, and thyroid hormone levels during pregnancy. *Environ Sci Technol*, 43,7909–15.

Ballester, F., Estarlich, M., Iniguez, C. et al. (2010). Air pollution exposure during pregnancy and reduced birth size: a prospective birth cohort study in Valencia, Spain. *Environ Health,* 9,6.

Fernandez, M.F., Olmos, B., Granada, A. et al. (2007). Human exposure to endocrine-disrupting chemicals and prenatal risk factors for cryptorchidism and hypospadias: a nested case-control study. *Environ Health Perspect*, 115(Suppl 1),8–14.

Freire, C., Ramos, R., Puertas, R. et al. (2010). Association of traffic-related air pollution with cognitive development in children. *J Epidemiol Community Health*, 64, 223–28.

Friguls, B., Garcia-Algar, O., Puig, C., Figueroa, C., Sunyer, J., Vall, O. (2009). Perinatal exposure to tobacco and respiratory and allergy symptoms in first years of life. *Arch Bronconeumol*, 45, 585–90.

Guxens, M., Ballester, F., Espada, M. et al. (2011). Cohort Profile: The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project. *International Journal of epidemiology*, 1-11.

Julvez, J., Ribas-Fito, N., Torrent, M., Forns, M., Garcia-Esteban, R., Sunyer, J. (2007). Maternal smoking habits and cognitive development of children at age 4 years in a population-based birth cohort. *Int J Epidemiol*, 36, 825–32.

Llop, S., Ballester, F., Estarlich, M., Esplugues, A., Rebagliato, M., Iniguez, C. (2010). Preterm birth and exposure to air pollutants during pregnancy. *Environ Res*, 110, 778–85.

Lopez-Espinosa, M.J., Vizcaino, E., Murcia, M. et al. (2010). Prenatal exposure to organochlorine compounds and neonatal thyroid stimulating hormone levels. *J Expo Sci Environ Epidemiol*, 20,579–88.

Lopez-Espinosa, M.J., Vizcaino, E., Murcia, M. et al. (2009). Association between thyroid hormone levels and 4,4'-DDE concentrations in pregnant women (Valencia,Spain). *Environ Res*, 109, 479–85.

Mendez, M.A., Torrent, M., Ferrer, C., Ribas-Fito, N., Sunyer, J. (2008). Maternal smoking very early in pregnancy is related to child overweight at age 5-7 y. *Am J Clin Nutr,* 87,1906–13.

Mendez, M.A., Garcia-Esteban, R., Guxens, M. et al. (2011). Prenatal organochlorine compound exposure, rapid weight gain and overweight in infancy. *Environ Health Perspect*, 119:272–78.

Polk, S., Sunye,r J., Munoz-Ortiz, L. et al. (2004). A prospective studyof Fel d1 and Der p1 exposure in infancy and childhood wheezing. *Am J Respir Crit Care Med*, 170, 273–78.

Ribas-Fito, N., Sala-Serra, M., Cardo, E. et al. (2002). Association of hexachlorobenzene and other organochlorine compounds with anthropometric measures at birth. *Pediatr Res*, 52,163–67.

Sunyer, J., Garcia-Esteban, R, Alvarez, M. et al. (2010). DDE in mothers' blood during pregnancy and lower respiratory tract infections in their infants. *Epidemiology*, 21, 729–35.

Torrent, M., Sunyer, J., Munoz, L. et al. (2006). Early-life domestic aeroallergen exposure and IgE sensitization at age 4 years. *J Allergy Clin Immunol*, 118, 742–48.

Torrent, M., Sunyer, J., Garcia, R. et al. (2007). Early-life allergen exposure and atopy, asthma, and wheeze up to 6 years of age. *Am J Respir Crit Care Med*, 176, 446–53.

Vrijheid, M., Martinez, D., Forns, J. et al. (2010). Prenatal exposure to cell phone use and neurodevelopment at 14 months. *Epidemiology*, 21, 259–62.

ESTUDIO EMPÍRICO DEL USO PROBLEMÁTICO DE LAS TECNOLOGÍAS DE ENTRETENIMIENTO DE LOS ADOLESCENTES ESPAÑOLES

Olatz López Fernández, Maria Luisa Honrubia Serrano y Montserrat Freixa Blanxart
Universidad de Barcelona
Correo electrónico: olatzlopez@ub.edu

Resumen

Esta investigación se centra en el uso problemático (o adictivo) de la videoconsola, el Internet y el teléfono móvil por parte de los adolescentes españoles. El propósito es determinar la existencia de uso habitual, en riesgo o problemático de estas tecnologías de entretenimiento para detectar en cada una su prevalencia en nuestra población. Este tema emergente tiene relevancia a nivel de las ciencias sociales y del comportamiento, puesto que desde hace una década se ha iniciado el estudio de la posible adicción a este tipo de entretenimientos derivado de las investigaciones en el área de la ludopatía en adolescentes. En este momento, la APA acaba de descartar la adicción al videojuego como trastorno a incluir en el próximo DSM-V, pero sigue estudiando la posibilidad de incluir la adicción al Internet. Paralelamente, diversos estudios epidemiológicos iniciados a partir de 2004 han determinado la prevalencia de la adicción al Internet y a la videoconosola en población joven. Pero pocos trabajos se centran en la etapa adolescente y las primeras escalas están empezando a ser difundidas en base a los criterios clínicos que se establecen para este tipo de problemática, tanto en países occidentales como sobretodo orientales. Nuestro estudio ha seleccionado una escalas validada para cada tecnología, que ha sido adaptada en caso necesario a la población adolescente española y se han incluido en un cuestionario ad hoc aplicado a una muestra de 1132 estudiantes de secundaria de Barcelona, de entre 12 y 18 años de edad entre 2008-2009. Se exponen los resultados preliminares hallados de forma exploratoria.

Los entretenimientos tecnológicos se han consolidado como una de las principales formas de ocio de la adolescencia. En España es una realidad constatada según la Asociación Española de Distribuidores y Editoras de Software y el Observatorio de la Producción Audiovisual, como uno de los principales juegos consumidos por este sector de la población, especialmente los videojuegos de consola, ordenador y por Internet, y recientemente a través del móvil.

La alerta social de adicción comportamental hacia este tipo de entretenimientos ha surgido en el ámbito científico desde aproximadamente 1995 en el caso de

Internet, 2000 en la videoconsola y 2005 en el móvil. La literatura científica aborda el tema de la adicción a las tecnologías de entretenimiento desde hace unos quince años (Griffiths & Hunt, 1995; Young, 1996), dado que la entidad clínica existe pero todavía no ha sido reconocida por ningún organismo oficial, ni la OMS ni la APA, aunque en el caso de la «adicción al internet» se está valorando su inclusión en el próximo DSM-V (la «adicción videoconsola» quedó descartado alrededor de 2008 por la falta de estudio epidemiológicos y la «adicción al móvil» aún es demasiado reciente para que haya estudios empíricos sobre la misma). En los tres casos de «adicción» a estas tecnologías, existe literatura científica tanto internacional como nacional, dado que este es uno de los países europeos con mayor presencia de casos clínicos de uso problemático o patológico en las tres tecnologías de entretenimiento, especialmente en los sectores de los adolescentes y de los jóvenes.

En referencia a los tipos de patrones de uso del juego, la literatura de la ludopatía determina tres principales tipologías de juego sin apuestas (el *playing*): el patrón de uso social (o habitual o controlado), que haría referencia al comportamiento de juego ocasional o regular con tecnologías, el patrón de uso de riesgo (o problemático), que se centra en cierta dependencia hacia la tecnología con una frecuencia prácticamente diaria que afectaría en la dedicación a otras actividades cotidianas (deberes, tareas del hogar, etc.) y/o relaciones con el entorno, demostrando dificultad por el auto-control; el patrón de uso problemático (o patológico), que mostraría una dependencia emocional con pérdida de control que interferiría con su vida cotidiana (amigos, familiares, y hábitos como la comida, el dormir, etc.).

El propósito de este trabajo es determinar la existencia de uso habitual, en riesgo o problemático de las tres tecnologías de entretenimiento para detectar su prevalencia de uso problemático la población adolescente de Barcelona.

MÉTODO

Participantes

La muestra estaba formada por 1132 adolescentes de la ciudad de Barcelona, estudiantes de centros de secundaria públicos y privados; con edades comprendidas entre 12 y 18 años (con una M = 14,55; DT = 1,816). En este punto, hay que señalar también que previamente se los pidió su consentimiento, donde se los garantía la confidencialidad y el anonimato de las respuestas, que los datos obtenidos serían sólo utilizadas para la presente investigación universitaria y que fuesen honestos en todo cuanto comunicasen. Al finalizar, se les agradeció su participación. Para participar en el estudio no era un requisito estar familiarizado con las tecnologías de entretenimiento, pues se trataba de hacer una exploración sobre su uso en la adolescencia, por el que sólo se limitó el intervalo de edad entre los 12 y 18 años.

Instrumentos

Para este estudio piloto de carácter cuantitativo, se diseñó un cuestionario *ad hoc*, con los siguientes bloques de contenido: (1) variables socio-demográficas (2) videojuegos de consolas; (3) videojuegos de ordenador y otros entretenimientos de Internet; (4) teléfono móvil. Junto con tres escalas:

PIESA - Internet: escala propia validada (López, Freixa y Honrubia, in progress)

PVP – Videoconsola (Tejeiro y Bersabé, 2002)

MPPUS – Teléfono móvil (Bianchi y Phillips, 2005).

Todas ellas, con el permiso de los autores, fueron adaptadas al castellano mediante el método de traducción-retro traducción para la población adolescente, excepto la PVP que fue facilitada por sus autores originales.

Procedimiento

Se solicitó permiso a los directores de los centros y estudiantes. Se realizó la administración del cuestionario durante una hora de clase, en que la investigador principal (primera autora del presente trabajo) estaba en el aula con los estudiantes para facilitar cualquier ayuda (resolución de dudas). El cuestionario era auto-aplicado y se realizó individualmente en silencio en grupo. En síntesis:

- Administración en IES de BCN (previo permiso directores)
- Las investigadoras sin los docentes en clase (solo investigadora)
- Se garantizaba el anonimato (confidencialidad)
- Era voluntario (el 100% participó)
- 1 hora de duración aproximadamente
- Durante el curso académico 2008-2009

Posteriormente, se realizó un análisis estadístico a nivel descriptivo de la muestra en relación a los tres entretenimientos tecnológicos: videojuegos de consuela, juegos de ordenador y webs y, del teléfono móvil. Se realizó también pruebas de carácter inferencial para determinar la relación que las variables de las tecnologías podían tener con el género y la edad, variables socio-demográficas principales que eran equilibradas en la muestra, así se obtuvieron los primeros resultados preliminares con SPSS (v. 15).

Resultados

En resumen, se procede a enumerar los resultados principales del estudio:

(1) Respecto al USO de estas tecnologías:

La mayoría tienen las tres tecnologías (ver figura 1): ordenador con internet, videoconsola y móvil (según porcentaje de mayor a menor).

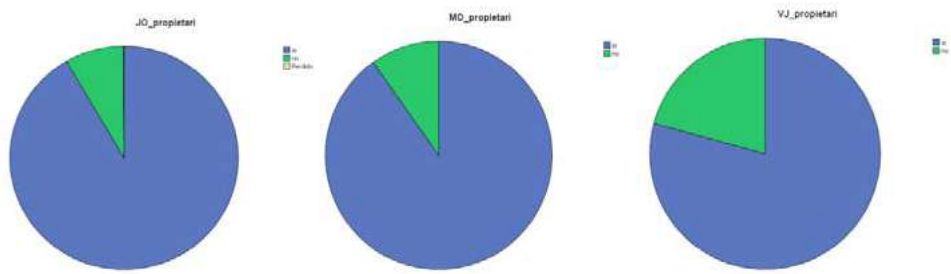


Figura 1. Descripción univariable del porcentaje de propietarios de ordenador con internet, de móvil y de videoconsola

Su grado de experto auto-percibido se encuentra entre el medio y bastante experto en su uso (ver figura 2).

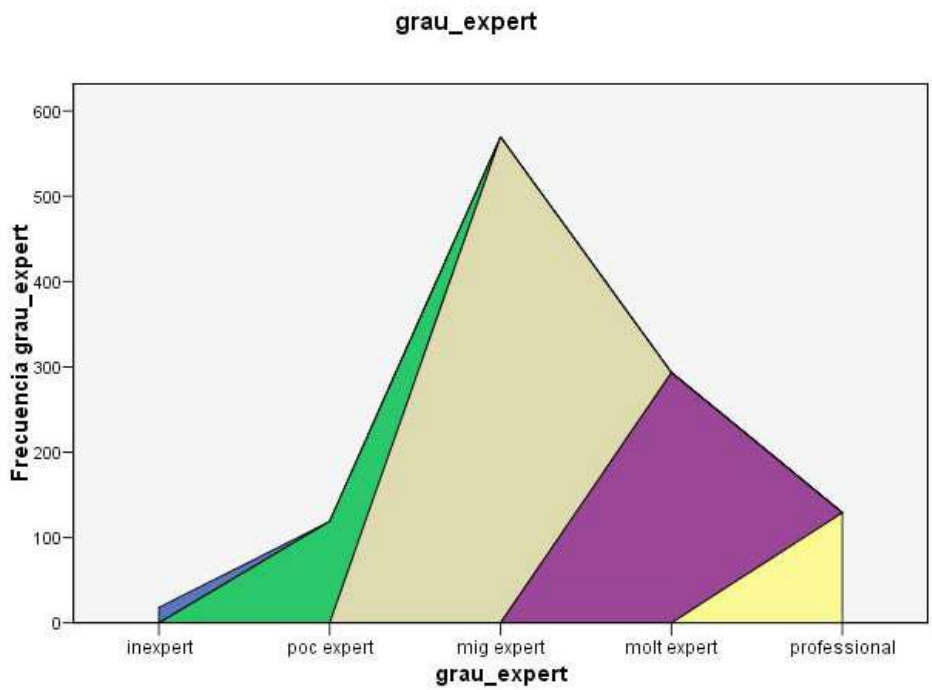


Figura 2. Descripción univariable del grado de experto percibido por los adolescentes respecto al uso que hacen de las tecnologías de entretenimiento

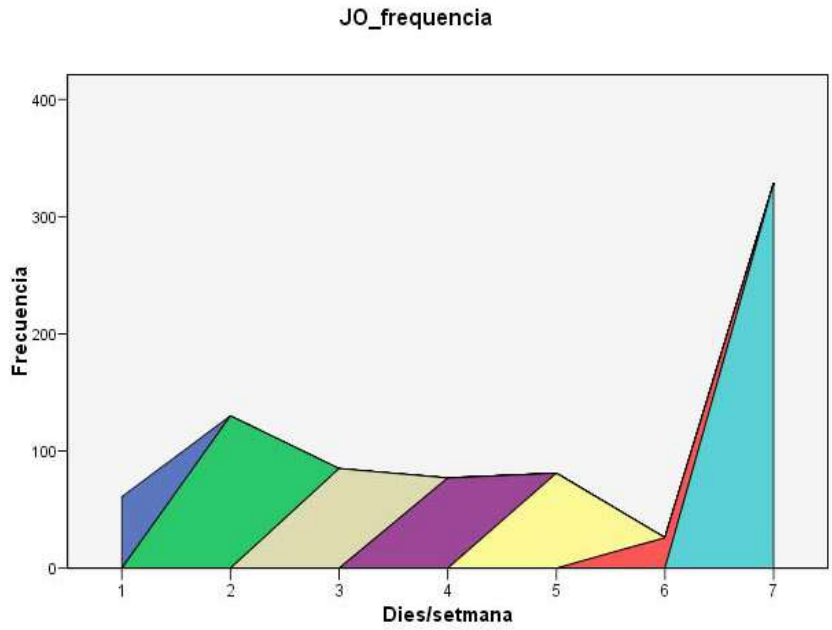


Figura 3. Descripción univariable la frecuencia de uso habitual de la videoconsola

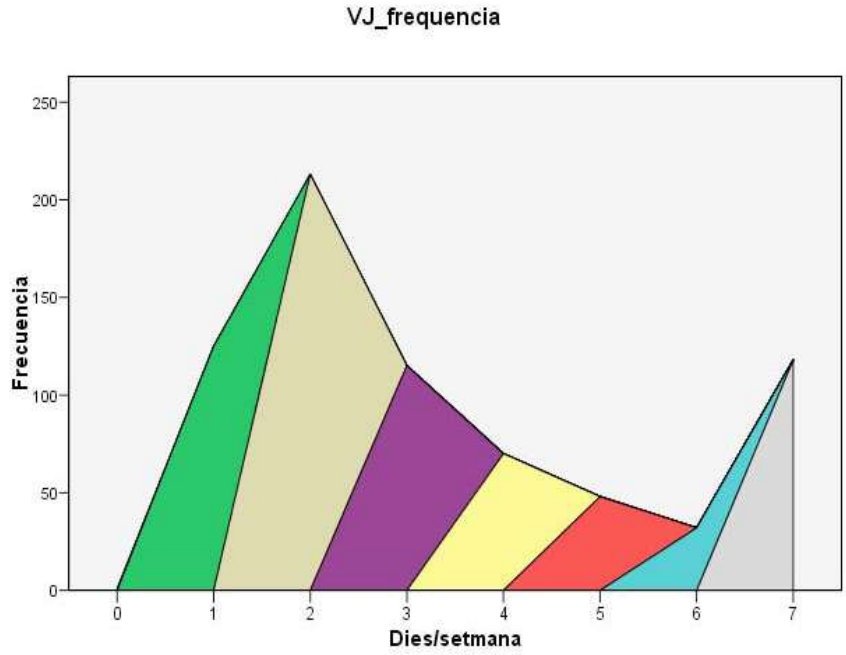


Figura 4. Descripción univariable la frecuencia de uso habitual del Internet

(2) En cuanto al TIEMPO DE USO de estas tecnologías:

Videoconsola: un 16% juega cada día, con una duración mínima de M=81,57 min., DT=83,761 y máxima de M=194,86, DT=171,186 (ver figura 3), por tanto serían usuarios entre el perfil en riesgo y el problemático.

Internet: un 41,6% se conecta cada día, con un tiempo mínimo de M=94,07 min., DT=80,925 y máximo de M=196,20, DT=156,329 (ver figura 4), por tanto éstos serían usuarios también entre el perfil en riesgo y el problemático.

(3) Referente a TIPOS DE USO (habitual, en riesgo, problemático):

Todavía estudiamos los puntos de corte para establecer las tres categorías de usuarios (como en Ludopatía). No obstante, en todas las tecnologías aparece un valor de prevalencia de usuarios problemáticos. Parece haber un uso excesivo (en cuanto a horas de dedicación y frecuencia de juego) entre el 15-40% de nuestros adolescentes y un uso problemático o patológico alrededor del 5-15% entre las tres tecnologías de entretenimiento.

CONCLUSIONES

Este estudio sigue siendo uno de los pocos realizados para determinar los patrones de uso problemático de estas tecnologías de entretenimiento por parte de los adolescentes. En este momento, se han perfeccionado los instrumentos y aplicado de forma masiva en tanto en Barcelona como en Londres, como ciudades de dos países europeos que se encuentran entre los primeros en adicciones comportamentales. Ello nos va a permitir obtener datos de mayor relevancia respecto a este uso y posible uso problemático de la videoconsola, Internet y Teléfono Móvil tanto en España como en Reino Unido, para tener una aproximación a nivel europeo de esta problemática de carácter psicológico que presentan algunos adolescentes de nuestro entorno.

REFERENCIAS

Bianchi, A. & Phillips, J.G. (2005). Psychological predictors of problem mobile phone use. *Cyberpsychology & behaviour, 8*(1) 39-51.

Freixa, M., López, O., y Honrubia, M.L. (2010). Construcción de un instrumento para medir el uso problemático de las tecnologías de entretenimiento adolescente. *VII Congreso Iberoamericano de Psicología,* Oviedo, 21 de julio de 2010.

Griffiths, M.D. & Hunt, N. (1995). Computer game playing in adolescence. Prevalence and demographic indicators. *Journal of Community and Applied Social Psychology, 5,* 189–193.

Tejeiro, R.A. & Besarbé, R.M. (2002). Measuring problem video game playing in adolescents. *Addiction, 97*, 1601-1606.

Young, K.S. (1998). Intenet addiction: the emergence of a new clinical disorder. *Cyberpsychology and Behaviour, 1,* 237-244.

ESTUDIO DE LA PERCEPCIÓN DEL RIESGO EN EL SECTOR DE LA CONSTRUCCIÓN: INVESTIGACIÓN HÍBRIDA DEL COMPORTAMIENTO DE RIESGO DE LOS TRABAJADORES DE UNA PLANTA DE PREFABRICADOS DE PIEZAS DE HORMIGÓN

Elisenda López Fernández[1] y Olatz López Fernández[2]
[1] Ministerio de Trabajo e Inmigración
[2] Universidad de Barcelona
Correo electrónico: elopezfe@mtin.es

Resumen

En España la construcción sigue siendo uno de los sectores de mayor siniestralidad laboral, por lo que se requiere del estudio continuado de los riesgos específicos de los trabajadores en la obra según su sector, puesto y rol desempeñado. El marco jurídico establece el deber del empresario de garantizar una protección eficaz en seguridad y salud laboral, donde se incluye la obligación de evaluar los riesgos en los puestos de trabajo mediante métodos objetivos que permitan identificar y valorar dichos riesgos. Este estudio centrado en la prevención de los riesgos laborales (PRL) del personal de producción de la obra de la línea 9 del metro de Barcelona, mediante los resultados extraídos con una combinación de técnicas cuantitativas propias del sector y cualitativas de carácter observacional (registro de conductas analizadas con GSEQ 5) y narrativo (entrevistas analizadas con Atlas-ti 5), los resultados de las cuales fueron trianguladas para validar el comportamiento de los trabajadores en la obra respecto a su comportamiento en la PRL de su puesto de trabajo. Los resultados muestran que se puede utilizar esta metodología en el estudio de PRL puede ser extensible para futuras investigaciones en el sector de la construcción, así como ofrecer posibles indicadores psicosociales a considerar en la elaboración de instrumentos que pretendan registrar el comportamiento laboral de los trabajadores de este sector.

Desde el campo de la seguridad y salud en el trabajo el estudio de los accidentes de trabajo ha sido tradicionalmente considerado sólo desde la ingeniería haciendo hincapié en los factores técnicos (diseño de equipos, procesos industriales, utilización de productos peligrosos, etc.) y los profesionales de este ámbito se han centrado en reducir los accidentes incidiendo en estos aspectos. Pero desde hace unas décadas se ha ampliado el campo de estudio a otras disciplinas, con métodos y técnicas que complementan a las anteriores y que, además, giran sobre el estudio del factor humano, desde una perspectiva científica, no sólo analizando al individuo (trabajador) su conducta, actitud hacia la prevención y el riesgo, sino también

la organización en sí, las relaciones interpersonales, la organización del tiempo de trabajo, etc. Este trabajo pretende ofrecer la posibilidad de incidir y estudiar en otros aspectos relacionados con el factor humano, en el ámbito de la prevención y, en concreto, aplicado a una obra de la construcción de la línea 9 del metro de Barcelona mediante la metodología híbrida.

Esta investigación abordó tres aspectos de los denominados Factores Humanos en la organización del trabajo y que la literatura científica asocia como posibles causas de accidentes de trabajo. La selección de estos tres factores aplicados al sector de la Construcción fue motivada principalmente porque este sector continúa siendo en la actualidad uno de los que tiene un elevado índice de incidencia de accidentes de trabajo, debido principalmente a sus peculiaridades: alta rotación de los trabajadores, condiciones extremas de trabajo, temporalidad, inmigración, trabajo a destajo, elevados ritmos de trabajo, condiciones climatológicas, sobrecarga física, etc.

Los principales objetivos son:

- Contribuir a la inclusión en el estudio y evaluación de los riesgos laborales de la perspectiva científico psicosocial de las condiciones de trabajo, en el estudio de las causas de estos accidentes en la Construcción.
- Establecer una relación entre las actitudes, la percepción de los riesgos y las conductas seguras/inseguras en el trabajo del personal de la Planta.
- Estudio objetivo del riesgo a través de la evaluación de riesgos laborales presentes y un análisis estadístico de los accidentes ocurridos en la Planta durante el periodo de estudio.

MÉTODO

Contexto

El contexto de esta investigación es un tajo, tramo, de una de las fases de construcción de la Línea 9 del Metro de Barcelona. Es una obra subterránea excavada con tecnología de tuneladora que a lo largo de su recorrido tiene varias estaciones, pozos de ataque, cocheras y una planta de prefabricado de piezas hormigonadas, dónde se fabrican las piezas dovelas, piezas de revestimiento del túnel que le dan consistencia y sostenimiento al mismo. Es una de las obras más emblemáticas y extensas que se han realizado en Catalunya a lo largo de estos últimos años.

Participantes

Trabajadores, encargados del tajo, mandos intermedios de la obra. La muestra eran todos los trabajadores de los dos turnos de la planta (mañana y tarde), 24 participantes, 20 trabajadores y 4 encargados y mandos. Todos eran hombres, la gran mayoría estaban situados en el rango de edad de entre 30-41 años (35%), en gene-

ral habían cursado estudios de primaria (70%) y llevaban trabajando en este Sector entre 0-10 años (50%)

Instrumentos

En el estudio de campo se emplearon varias metodologías para evaluar los riesgos, integrando la perspectiva técnica de la prevención de riesgos (métodos jurídico-técnicos) y técnicas psicosociales, entre las que se encuentran:

- Para el estudio de la percepción del riesgo, una escala de actitudes basada en el método de evaluación EDRP-T
- Para el análisis de conductas seguras/inseguras se utilizó metodología observacional, se elaboraron instrumentos de recogida de datos tipo checklists «ad hoc» para esta investigación y fotografías y se empleó el SDIS-GSEQ para su análisis
- Para el estudio de las actitudes seguras se administró la Escala Cyclops
- Completándose esta información con entrevistas semiestructuradas al personal encargado, trabajadores, mandos intermedios, responsables de seguridad y salud en la obra y posteriormente analizando el contenido cualitativo con el CAQDAS ATLAS.ti
- Análisis estadístico de datos de accidentes de trabajo facilitados por el Servicio de Prevención de la UTE y por el Servicio Médico de la obra.
- Evaluación objetiva de los riesgos laborales presentes en el tajo con metodología de la Generalitat de Catalunya.

Procedimiento

Se administraron todos los instrumentos de recogida de datos cuantitativos y cualitativos a la vez, es decir, que se aplicó un diseño híbrido de implantación simultánea de distinta prioridad, con énfasis en los métodos cuantitativos, es decir, con un diseño híbrido CUAN+cual (según Johnson y Onwuegbuzie, 2004).

RESULTADOS

Esta investigación está enmarcada dentro de una tesis doctoral desarrollada en la Universitat Politecnica de Catalunya (UPC) en el Departamento de Organización de Empresas, cuyos resultados publica la Web de la misma Universidad. Sin embargo, se muestran algunos relacionados con las características de la muestra (véase figura 1), con la evaluación EDRP-T (véase figura 2), con los *checklist*, con las observaciones (véase figura 3), con la escala cylcops (véase figura 4) y las entrevistas (véase figura5).

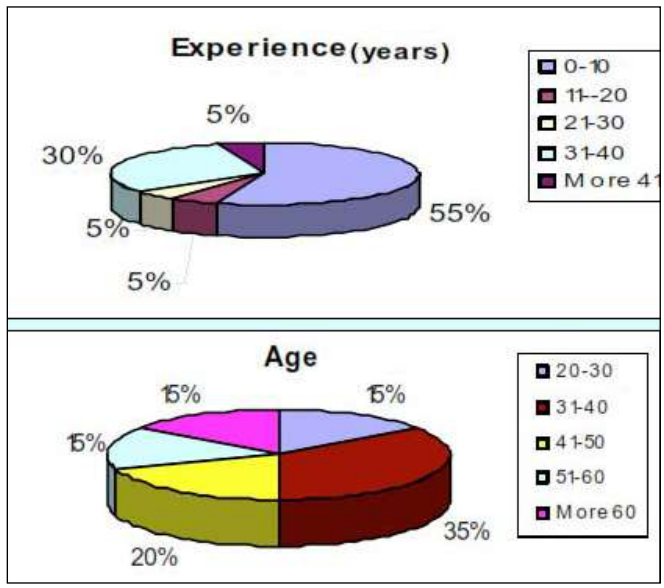


Figura 1. Resultados de la muestra (cuantitativos)

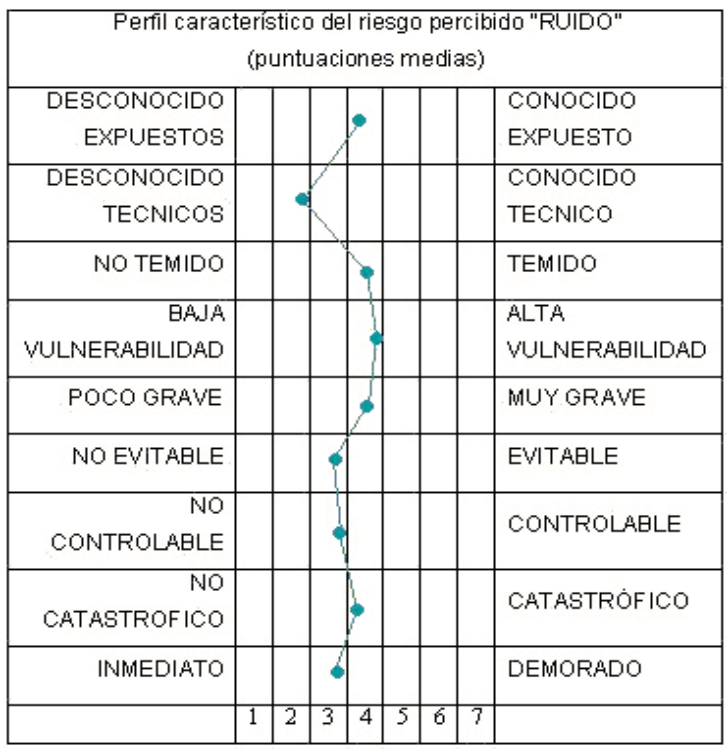


Figura 2. Resultados del cuestionario EDRP-T (cuantitativos)

Dados Condicionados
A1 A2 A3 A4 A5 A6 A7 A8

F1| 0.65: -0.88: 0.19: -0.60: -0.42: -0.73: -0.51: -0.30:|
F2| -0.65 0.88 -0.19 0.60: 0.42: 0.73: 0.51: 0.30:|

Dados Condicionados
A9 A10 A11 B1 B2 B3 B4 B5

F1| 0.00: 0.00: -0.42: 3.39: -0.79: -0.30: -0.73: -0.60:|
F2| 0.00: 0.00: 0.42: -3.39: 0.79: 0.30: 0.73: 0.60:|

Dados Condicionados
B6 B7 B8 B9 B10 B11 B12 C1

F1| -0.30: 0.62: -1.09: 2.31: -0.42: 1.00: 2.03: 0.00:|
F2| 0.30: -0.62: 1.09: -2.31: 0.42: -1.00: -2.03: 0.00:|

Dados Condicionados
C2 C3 C4 C5 C6 C7 C8 D1

F1| 0.00: 0.00: -0.30: -0.30: -0.30: 0.00: 0.00: 0.00:|
F2| 0.00: 0.00: 0.30: 0.30: 0.30: 0.00: 0.00: 0.00:|

Dados Condicionados
D2 E1 E2

F1| 0.00: 0.00: -0.67:|
F2| 0.00: 0.00: 0.67:|

Figura 3. Resultados del análisis retrospectivo con retardo 0 (cuantitativos)

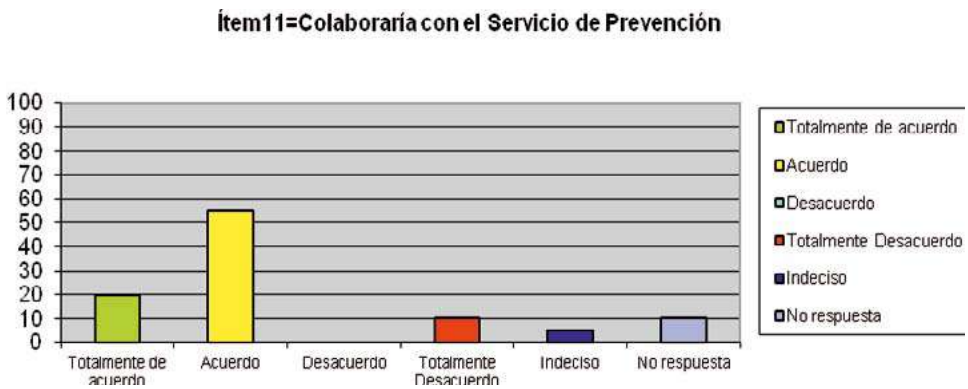


Figura 4. Resultados de la Escala Cyclops Actitudes hacia la Seguridad en el Trabajo (cuantitativos)

Participante 1. Debería informarse a los trabajadores sobre las consecuencias sobre la salud de los trabajadores cuando se manipula éste tipo de productos. Saber hasta qué punto se puede tocar y los equipos de protección más adecuados para ello, que te protejan.

Participante 2**. SABER LOS TÓXICOS CON LOS QUE ESTAMOS TRABAJANDO.**

Figura 5. Resultados de la entrevista a los trabajadores de las dovelas (cualitativos)

En síntesis, el resultado cualitativo se sintetiza en una serie de redes semánticas relacionadas con los objetivos a lograr mediante el análisis de datos cualitativos en base a las entrevistas semi-estructuradas (véase figura 6).

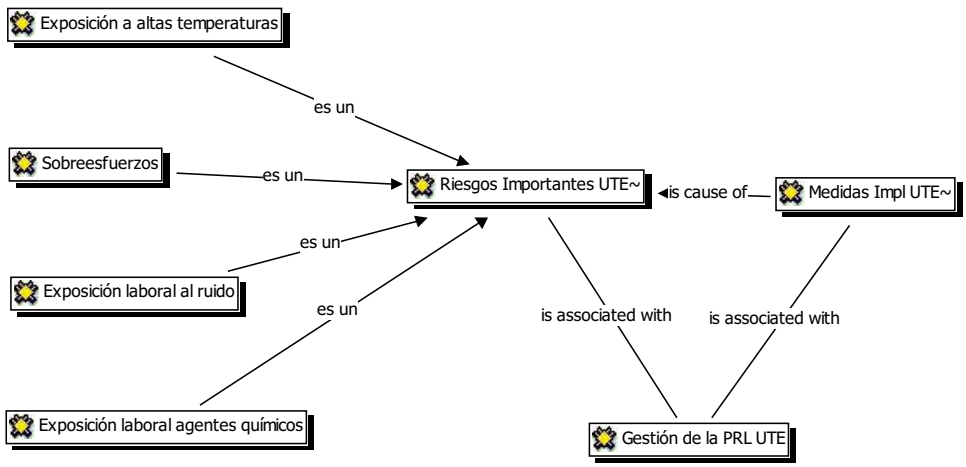


Figura 6. Resultados la parte cualitativa: Network «riesgos en la planta de dovelas»

En cambio, el cuantitativo una serie de análisis longitudinales respecto a la evolución de los accidentes durante el periodo del estudio de campo (2004-2005) (véase figura 7) y los extraídos con el método de la Generalitat de Cataluña (véase figura 8).

Figura 7. Resultados parte cuantitativa: Evolución accidentes laborales en la UTE L9

Ficha número 3	IDENTIFICACIÓN GENERAL DE RIESGOS
DATOS IDENTIFICATIVOS DE LA EMPRESA: UTE XX	DATOS DE LA EVALUACIÓN
RAZÓN SOCIAL: UTE XX CNAE: construcción ACTIVIDAD: Construcción CCC:	Duración: Un mes Tipo: Inusual
DIRECCIÓN CT: Anselm de Riu, 8 CP:08924 LOCALIDAD: Santa Coloma de Gramanet TEL: 93 390 55 00	Realización: Planta dovelas Construcción camillas

núm.	PUESTO DE TRABAJO	FORMA DEL ACCIDENTE																								TIPO MT				S.ESP.		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	SE	MA	ME
1	Cargar y descargar maderos			X										X																			
2	Cortar maderas con la sierra circular de mesa								X		X							X															
3	Montaje la camilla a golpe de martillo y clavos							X		X				X																			
4	Transporte de camilla a losa de acopio con la carretilla elevadora					X							X											X									

CÓDIGOS DE UTILIZACIÓN

01. Caída de personas a distinto nivel
02. Caída de personas al mismo nivel
03. Caída de objetos por desplome
04. Caída de objetos por manipulación
05. Caída de objetos desprendidos
06. Pisadas sobre objetos
07. Choques contra objetos inmóviles
08. Contactos con elementos móviles de la máquina
09. Golpes por objetos o herramientas
10. Proyección de fragmentos o partículas
11. Atrapamientos por o entre objetos
12. Atrapamientos por vuelco de máquinas
13. Sobreesfuerzos
14. Exposición a temperaturas extremas (altas/bajas)
15. Contactos térmicos
16. Contactos eléctricos
17. Inhalación o ingestión de sustancias nocivas
18. Contactos con sustancias cáusticas / corrosivas
19. Exposición a radiaciones
20. Explosiones
21. Incendios
22. Causados por seres vivos
23. Atropellos, golpes y choques con o contra vehículos
24. Accidentes de tráfico
25. Causas naturales (infarto, embolia, etc.)
26. Otras
27. Enfermedades causadas por agentes químicos (dermatitis y alergias)
28. Enfermedades causadas por agentes físicos (vibraciones y ruido)
29. Enfermedades causadas por agentes biológicos
30. Enfermedades causadas por otras circunstancies

Figura 8. Resultados parte cuantitativa: Registro del Método Generalitat de Cataluña

CONCLUSIONES

En España no hay todavía suficientes estudios que aborden el análisis de las causas de los accidentes de trabajo, desde una perspectiva psicosocial, en el Sector de la Construcción «in situ», es decir. Investigación aplicada y real cuyos resultados puedan aportar y ayudar a mejorar la gestión cotidiana de los riesgos laborales de una empresa constructora dentro de la obra y, por tanto, a reducir los accidentes de trabajo.

En este trabajo las autoras proponen con esta investigación focalizar el estudio en tres factores psicosociales: la percepción del riesgo, las actitudes seguras y las conductas seguras/inseguras de los trabajadores. Se autorizó por parte de las constructoras de la obra a entrar en el recinto para estudiar, observar, analizar, fotografiar, administrar cuestionarios y otras pruebas, así como entrevistar al personal durante la jornada laboral, en ambos turnos y en horario de fin de semana, para saber su opinión sobre la prevención de riesgos laborales de la obra y conocer las condiciones de seguridad y salud en la que trabajan

En las últimas Encuestas Nacionales de Condiciones de Trabajo editadas por el Instituto Nacional de Seguridad e Higiene en el Trabajo (Ministerio de Trabajo e Inmigración) en el que pregunta a una muestra representativa de los trabajadores de todos los sectores su opinión acerca de varios aspectos relativos a la prevención de riesgos laborales, en uno de sus ítem, sobre cuál creían ellos que era la principal causa de accidentes de trabajo contestaron en primer lugar que era el «exceso de confianza». Este ejemplo denota que en la actualidad todavía, en el ámbito laboral, tenemos camino por recorrer para consolidar entre los trabajadores y empresarios una verdadera y real cultura preventiva.

REFERENCIAS

Johnson, B. y Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

EL EFECTO DE LA DEFINICIÓN DE FIJACIÓN OCULAR SOBRE RESULTADOS EXPERIMENTALES

Elisa Pérez- Moreno, Ángela Conchillo y Miguel Ángel Recarte
Universidad Complutense de Madrid
Email: elisaperez@psi.ucm.es

Resumen

Para aislar una fijación en el conjunto de datos del ojo y de la mirada se requiere unos criterios complejos, que además no son universalmente aceptados. Algunos autores han defendido la importancia de establecer dichos criterios demostrando que las características oculares, como la duración o el número de fijaciones, pueden verse drásticamente influenciadas por los mismos. Sin embargo, no se han encontrado estudios que demuestren cómo los criterios que definen una fijación ocular pueden afectar directamente a los resultados experimentales y a las conclusiones que se extraigan a partir de ellos. En esta investigación el conjunto de criterios de fijación empleado en un estudio previo con tareas de tipo verbal, espacial y de búsqueda visual (Pérez-Moreno, Conchillo, & Recarte, 2011) fue modificado hacia un conjunto de un criterios más restrictivos y hacia otro más laxos. Se demuestra que el criterio empleado modifica, no sólo las características oculares sino también el efecto que otras variables tienen sobre éstas.

Las fijaciones oculares han sido definidas como un momento relativamente estable del ojo durante el cual es posible el procesamiento de información exterior. Sin embargo, para aislar una fijación del conjunto de datos del ojo se requiere unos criterios y un esquema de partición que puede llegar a ser complejo y, en algunas ocasiones, arbitrario. Según Shebilske & Fisher (1983) no es suficiente con construir un aparato que permita determinar en cada momento dónde se encuentra el ojo, sino que también debe distinguir entre las posiciones del ojo medidas durante movimientos sacádicos y fijaciones oculares. Heller (1983) desarrolla una rutina de identificación de fijaciones que incluye dos criterios, por un lado, una tasa de supresión para las diferencias en la distancia espacial entre dos valores consecutivos y por otro, la duración mínima de una fijación. Sin embargo, no parece existir un acuerdo claro acerca de cuáles son los criterios que deben emplearse, ni los valores que deben adoptar para considerar un conjunto de datos dentro de una misma fijación.

Karsh & Breitenbach (1983) estudian el efecto de estos criterios sobre algunas características de las fijaciones, como el número o duración de éstas, demostrando cómo los resultados de estas variables se modifican drásticamente y resaltando la importancia de la definición precisa de fijación ocular.

No obstante, parece obvio que, modificando los criterios que definen qué es una fijación ocular, las características de las fijaciones oculares se vean alteradas. Por ejemplo, si se utiliza un criterio más restrictivo para incluir un dato nuevo dentro de una fijación (a partir de ahora, criterio conservador), parece razonable que, la duración de las fijaciones sea menor que si se utiliza un criterio más laxo (a partir de ahora, criterio liberal). Más interesante sería, por tanto, estudiar si los efectos de variables externas sobre los movimientos oculares y sus características se ven modulados por un cambio en la definición de fijación ocular. En este sentido, no se han encontrado estudios que demuestren cómo los criterios de fijación ocular pueden afectar directamente a los resultados experimentales y, por tanto, a las inferencias y conclusiones que se extraen a partir de ellos. Salvucci & Goldberg (2000) proponen una clasificación de algoritmos y los comparan de acuerdo a unas características cualitativas, sin embargo, los propios autores consideran que el siguiente paso es evaluar y comparar los algoritmos con medidas cuantitativas y examinar cómo los algoritmos pueden afectar a los análisis posteriores en distintos contextos y aplicaciones.

Nuestro objetivo, por tanto, es comprobar el efecto de la definición de fijación ocular no sólo sobre características oculares (duración de las fijaciones, variabilidad horizontal y vertical y amplitud sacádica) sino también sobre variables respuesta (proporción de estímulos target mirados y proporción de estímulos target detectados entre aquellos que fueron previamente mirados) de una investigación previa con tareas de tipo verbal, espacial y de búsqueda visual (Pérez-Moreno et al., 2011) y comprobar si los resultados y conclusiones que fueron extraídos de la investigación original pueden mantenerse después de modificar la definición de fijación ocular o si por el contrario, se ven alterados por los nuevos criterios de fijación ocular adoptados.

MÉTODO

Participantes

Se han utilizado los datos oculares de 29 participantes de Pérez-Moreno et al., (2011), en adelante, investigación original.

Diseño del estudio original

El diseño del experimento original fue intrasujeto con una variable independiente (tipo de tarea cognitiva: control, verbal y espacial) y seis variables dependientes: cuatro variables oculares (duración de las fijaciones, variabilidad horizontal y vertical de las fijaciones y amplitud sacádica) y dos variables respuesta (proporción de estímulos target mirados y proporción de estímulos target detectados entre los que fueron mirados).

Las tareas que los participantes tenían que realizar fueron: una tarea de Búsqueda Visual (BV) que consistió en detectar un símbolo target que variaba en

orientación respecto a los distractores en un display dinámico. La BV se realizó como tarea simple y en situación de doble tarea, combinada simultáneamente con una tarea verbal y otra espacial. La tarea verbal (V) consistió en pronunciar palabras encadenadas, y la tarea espacial (E) que consistió en moverse mentalmente a través de una matriz de 3x3. Así pues las condiciones experimentales fueron tres: BV, BV+V, y BV+E.

A estas tres condiciones experimentales hay que añadir las tres condiciones que son objeto del presente estudio, proporcionadas por el Criterio para definir una fijación: Conservador, Original, Liberal.

Materiales y procedimiento

Los datos oculares fueron obtenidos mediante un sistema de registro ocular (ASL Model 5000) a 50 Hz, y en esta investigación se variaron los parámetros que definen una fijación ocular, utilizando para ello un criterio más conservador y otro más liberal que el empleado originalmente.

Tabla 1. Parámetros que definen una fijación ocular y valores concretos de estos parámetros en el estudio original y tras adoptar un criterio más conservador y otro más liberal que el original

Parámetros	Criterio Conservador	Criterio Original	Criterio Liberal
P1. Mínima muestra de datos (primera muestra)	6	5	4
P2. Dispersión de la primera muestra (en grados de a.v.)	< 0.5 D.T.	< 1 D.T.	< 1.5 D.T.
P3. Siguientes muestras de datos	2	3	4
P4. Dispersión de la siguientes muestras	<0.5 D.T.	<1 D.T.	< 1.5 D.T.
P5. Posición final de la fijación (excepto datos espurios)	<1 D.T.	< 1.5 D.T.	< 1.5 D.T.

Parámetros que definen una fijación ocular. Se parte de una muestra inicial de datos oculares (P1) cuya desviación típica (en grados de ángulo visual) no sea superior a P2. Las subsiguientes muestran se toman de P3 en P3. Para que una muestra sea incluida en la misma fijación que la inicial no debe alejarse más de P4 desviaciones típicas de la inicial. El cálculo de la posición final de la fijación será la media de todos los datos incluidos excepto si alguno de ellos se separa de la original más de P5 desviaciones típicas. Ver Tabla 1 para los valores concretos de estos parámetros, tanto en el estudio original, como adoptando un criterio más conservador y más liberal que éste.

RESULTADOS

La Tabla 2 muestra los resultados originales de un ANOVA de medidas repetidas por tipo de tarea, comparados con los obtenidos en este estudio utilizando un criterio más conservador y otro, más liberal que el del estudio original.

Tabla 2. Medias y error típico de la media de las variables duración (DUR), variabilidad horizontal (VAR X) y vertical (VAR Y) de las fijaciones, amplitud sacádica (AMP), proporción de estímulos mirados (PEM) y proporción de estímulos detectados entre los que fueron mirados (PED/PEM)*

		Criterio conservador				Criterio original				Criterio liberal			
	Tarea	**Mean (SE)**	**p**	**η^2**	**1-β**	**Mean (SE)**	**p**	**η^2**	**1-β**	**Mean (SE)**	**p**	**η^2**	**1-β**
DUR			.549				.071				.032*	.12	.65
	BV	140(3)				184(6)				179(6)			
	BV + V	137(3)				178(7)				181(9)			
	BV + E	138(3)				190(7)				203(11)			
Var X			<.001**	.27	.98		<.001**	.15	.8		.11		
	BV	49(3)				47(1)				47(1)			
	BV + V	39(2)				41(2)				48(3)			
	BV + E	38(2)				40(2)				43(2)			
Var Y			<.001**	.48	1		<.001**	.3	.99		.041*	.11	.61
	BV	35(2)				36(2)				36(1)			
	BV + V	28(2)				31(2)				32(1)			
	BV + E	27(1)				30(1)				33(2)			
AMP			<.001**	.61	1		<.001**	.46	1		<.001**	.32	.99
	BV	6.48(0.32)				6.29(0.43)				6.45(0.44)			
	BV + V	4.9(0.27)				5.28(0.25)				5.86(0.27)			
	BV + E	4.12(0.23)				4.28(0.23)				4.92(0.23)			
PEM			.004**	.18	.89		.002**	.2	.9		.098		
	BV	.64(.05)				.75(.02)				.57(.05)			
	BV + V	.47(.05)				.64(.02)				.47(.04)			
	BV + E	.47(.04)				.59(.02)				.44(.04)			
PED/ PEM			.001**	.21	.93		.001**	.38	.99		.332		
	BV	.29(.04)				.41(.03)				.22(.03)			
	BV + V	.17(.03)				.17(.03)				.15(.04)			
	BV + E	.13(.03)				.29(.03)				.20(.04)			

* Los datos originales son comparados con los obtenidos después de filtrar los datos con un criterio de definición de fijación ocular conservador y otro, liberal.

Los resultados muestran que si el criterio utilizado es más conservador que el original, se encuentran diferencias estadísticamente significativas para las variables variabilidad horizontal y vertical de las fijaciones, amplitud sacádica y proporción de estímulos mirados y detectados entre los previamente mirados por el tipo de tarea cognitiva que se esté llevando a cabo, al igual que en el estudio original. Sin embargo, si el criterio es más liberal que el original, no sólo desaparecen la

mayor parte de estas diferencias, sino que aparecen diferencias estadísticamente significativas en la variable duración de las fijaciones según la tarea cognitiva que se esté realizando.

Además, para comparar el criterio conservador con el liberal, se realizó un ANOVA de medidas repetidas tarea cognitiva x criterio (3 x 2) para cada una de las variables dependientes.

Como era de esperar, para todas las variables oculares es significativo el efecto del *criterio*, en el sentido de que con un criterio conservador la duración de las fijaciones es menor, con F (1,28) = 95.55, p <.001, η^2 = .77, la variabilidad horizontal y vertical es menor, con F (1,28) = 4.78, p = .037, η^2 = .146 y F (1,28) = 9.89, p = .004, η^2 = .261, respectivamente, y la amplitud sacádica es menor, con F (1,28) = 9.56, p = .004, η^2 = .255, que cuando se utiliza un criterio más liberal.

Más interesante es comprobar el efecto diferencial del criterio sobre los niveles de la variable tarea cognitiva, tal y como indican las interacciones *criterio x tarea*:

Para la variable duración de las fijaciones la interacción es significativa con F (2,56) = 6.31, p = .003, η^2 = .184, en el sentido que el aumento en la duración de las fijaciones al utilizar un criterio más liberal es más pronunciado al realizar BV + tarea cognitiva (doble tarea) que para BV sola (tarea simple), y al realizar BV + E frente a BV + V.

Para la variable variabilidad horizontal la interacción es significativa con F (2,56) = 8.74, p < .001, η^2 = .238, en el sentido que el aumento en la variabilidad de las fijaciones utilizando un criterio liberal es mayor cuando se realiza BV + tarea cognitiva que para BV sola, y al realizar BV + V frente a BV + E.

Para la variable variabilidad vertical la interacción es significativa con F (2,56)=3.96, p = .02, η^2 = .124, en el mismo sentido que para la variabilidad horizontal.

Para la variable amplitud sacádica la interacción es significativa con F (2,56) = 5.69, p = .006, η^2 = .169, en el sentido que la disminución en la amplitud sacádica al utilizar un criterio liberal (frente a uno conservador) es más pronunciada cuando se realizar BV + tarea cognitiva que para BV sola.

CONCLUSIONES

Según los resultados obtenidos, el criterio utilizado modifica, no sólo los valores absolutos de las características oculares, tal y como anteriormente habían concluido otros autores (Karsh & Breitenbach, 1983; Manor & Gordon, 2003) sino que también modifica el efecto que otras variables tienen sobre éstas, obteniéndose conclusiones distintas para unos mismos datos oculares como origen.

Nuestros resultados muestran, como era de esperar, que al utilizar un criterio más conservador disminuyen la duración de las fijaciones, la variabilidad horizon-

tal y vertical de éstas y la amplitud sacádica frente a utilizar un criterio más liberal. Sin embargo, estos resultados llegan más allá indicando que los efectos que produce el criterio sobre las variables oculares son más pronunciados cuando se realiza una tarea doble frente a una simple (tarea cognitiva más tarea de búsqueda visual frente la búsqueda visual sola), tal y como indican las interacciones de la variable criterio con la variable tarea cognitiva. Es decir, se producen efectos diferenciales según las condiciones experimentales.

Otros resultados novedosos encontrados en este estudio indican que, mientras que se utiliza un criterio conservador las diferencias encontradas en las variables oculares y respuesta en el estudio original por la variable tarea cognitiva (BV, BV+V, BV+E) se mantienen, al cambiar a un criterio más liberal las diferencias desaparecen o aparecen nuevas diferencias anteriormente no encontradas.

A la vista de estos resultados, se muestra necesaria la creación de una definición precisa y universal de fijación ocular, ya que unos mismos datos oculares pueden dar lugar a distintas interpretaciones según la definición de fijación ocular que se esté utilizando.

REFERENCIAS

Heller (1983). Problems of on-line processing of EOG-Data in reading. In: R. Groner, C. Menz, D. Fisher and R. A. Monty (Eds.), *Eye Movements and Psychological Functions: International Views* (pp. 53–64). Hillsdale, NJ: Lawrence Erlbaum Associates.

Karsh, R., & Breitenbach, F. W. (1983). Looking at looking: The amorphous fixation measure. In: R. Groner, C. Menz, D. Fisher and R. A. Monty (Eds.), *Eye Movements and Psychological Functions: International Views* (pp. 53–64). Hillsdale, NJ: Lawrence Erlbaum Associates.

Manor, B.R., & Gordon E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of Neuroscience Methods, 128*, 85-93.

Perez- Moreno, E., Conchillo, A., & Recarte, M. A. (2011). Interference in visual perception by verbal and spatial cognitive activity. *The Spanish Journal of Psychology, 14* (2), 556- 568.

Salvucci, D.D., & Goldberg, J.H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the eye tracking research and application symposium* (pp. 71-78). NY: ACM Press.

Shebilske, W.L., & Fisher, D.F. (1983). Understanding extended discourse through the eyes: how and why. In: R. Groner, C. Menz, D. Fisher and R. A. Monty (Eds.), *Eye Movements and Psychological Functions: International Views* (pp. 53–64). Hillsdale, NJ: Lawrence Erlbaum Associates.

METHODOLOGICAL ISSUES OF NATURALISTIC DRIVING OBSERVATION WITH EQUIPPED CARS

Pedro Valero, Anita Tontsch, Ignacio Pareja y Mar Sanchez
Universidad de Valencia
Correo electrónico: pedro.valero-mora@uv.es

Abstract

The project PROLOGUE funded by the 7th Framework Programme of the European Commission aimed to explore the methodology of observing drivers while driving in a natural setting. Naturalistic driving observation includes the use of video cameras for capturing the driver and the environment of the car and the recording of vehicle parameters to obtain information about the drivers` behaviour while driving. This approach provides information that traditional methods like laboratory experiments (simulations, observational studies etc.) or traffic information do not enable. However, this method also has some metrological drawbacks like the low control of external variables, the issue of data storage and especially of data analysis. This paper will summarize our experience with the use of highly instrumented vehicle for field trials and the methodological issues we had to face.

Naturalistic driving observation means to observe road users` behaviour in an unobtrusive way, in their natural setting, over a long period of time. Drivers` behaviour comprises not only driving parameters but can also include in-car behaviour like eye glances, head- and body movements. For the observation a real car is equipped with certain measuring instruments and cameras that capture driving parameters, the driver himself as well as his environment. These instruments ranges from small devices that can be installed in the participants own cars to extensive equipment (Backer-Grøndahl et al. 2009; Valero-Mora et al., 2010).

The methodology of naturalistic driving (ND) has certain strength and weaknesses. To the advantages count that it enables a direct observation of behaviour as well as of critical traffic incidents, near-crashes and/or accidents. The unobtrusive and realistic observation is objective and results in a high ecological validity and eliminates disadvantages of the rather traditional methods like e.g. self-reports (biased by social desirability and driving simulation (artificial situation). Naturalistic driving furthermore supplies a big variety of data; besides the subjects` driving behaviour one moreover gathers information about road user interaction, the drivers` behaviour within the car, GPS information of driving path (map match-

ing), and as a consequence enables in-depth information about factors contributing to crashes/near-crashes (Backer-Grøndahl et al. 2009).

One weaknesses of ND is the low control one has over external variables. The driving observation is meant to be naturalistic, that is, there is no experimental setting or intervention. Another drawback is the resource demand that is bonded to this kind of observation; ND studies are meant to go over a long period of time what demands resources in terms of a sample, the data gathering, the storage and the analysis of a vast amount of data. Another negative point is the possibility of biased samples. As the drivers` daily driving is in the focus of observation, the ones that agree in participation might not represent a representative sample of all drivers but rather one of safer drivers (Backer-Grøndahl et al. 2009).

PROLOGUE, what stands for «PROmoting real Life Observation for Gaining Understanding of road user behaviour in Europe» has been a recently finished project under the 7th Framework Programme co-funded by the European Commission. The project`s aim was the reduction of road causalities in Europe by exploring, testing and developing the naturalistic driving methodology. Main objective of the project was to prove the feasibility and usefulness of a large-scale naturalistic driving observation study for road safety researchers and other stakeholders. Other project partners besides the University of Valencia (ES) where SWOV (NL), CERTH/HIT (GR), KfV (AT), Loughborough University (UK), Or Yarok (IL), TNO (NL), TØI (NO), Test&Training (AT).

In order to gain information about the methodology and to comply the project`s aims, small scale field trials with different characteristics have been conducted in several countries. The field trial in Valencia has been realized using a highly instrumented car called ARGOS which was funded and is owned by the DGT (stands for Dirección General de Tráfico, Madrid) in Spain and developed in collaboration with the UPM (stands for Universidad Politécnica de Madrid, Spain).

SMALL-SCALE FIELD TRIAL IN VALENCIA

ARGOS is a standard SEAT Alhambra (see Figure). The ARGOS car is equipped with a number of sensors recording several parameters that can be grouped in the following categories: dynamics of the car, driver vehicle interaction, comfort of the driver, indicators in the car, environmental conditions, data acquisition parameters for the driver and experimental events.

The car furthermore includes seven video cameras observing the driver and his environment. Figure 2 demonstrates the location of these cameras.

ARGOS had not been tested in practice before; it was not clear whether it would perform adequately in actual trials. Therefore, we set up a trial with the primary objective of exploring its functioning and its use as part of a naturalistic driving study. But in order to establish a realistic setting for testing the car, we also set as an objective of the trial to observe the effects of the use of Nomadic Devices

while driving. This second objective can be regarded as secondary/instrumental with respect to the main objective of evaluating the capabilities and limitations of the ARGOS car (Valero-Mora et al., 2010).

Figure 1. The ARGOS car; left the outside view and right the insight view of the equippment in the back

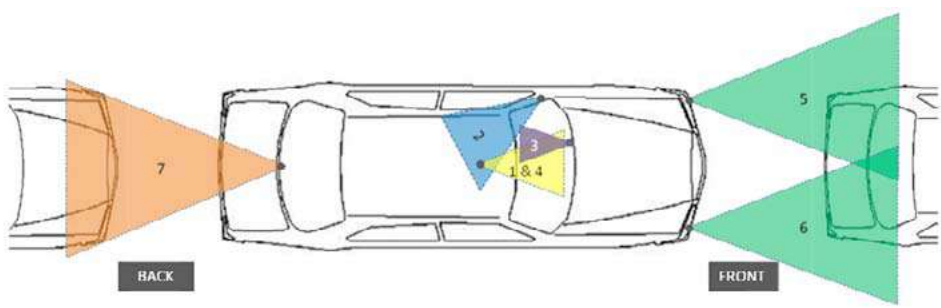


Figure 2. Camera positions in ARGOS

METHODOLOGICAL ISSUES OF THE TRIAL

In this section the methodological issues that occur during the trial while be explained. Some of them, especially technical ones may nowadays not be problematic anymore due to the further development of technology.

Data storage: While driving, video and sensor data was collected continuously. Due to the continuous data collection 72GB of data have been produced in about two hours of driving. Data had to be downloaded daily in order to avoid an overloading of the internal storage (Valero-Mora et al., 2010).

Setting of the trial: The field trial was conducted with five experienced drivers that drove the car for four consecutive days each. Every day each driver received a list of destinations that he has to go to one after another. It was by the driver`s decision to select the way for reaching this destinations. On two of the four days they

were free to use a navigation device or any other in-vehicle information system or nomadic device. The other two days he was not allowed to use these systems. The aim was to investigate whether the use of IVIS while driving influenced the participants' way to drive and to get information about their behaviour while paying attention to these secondary tasks. This experimental setting does not confirm with the concept of naturalistic driving, however, considering the small extent of this small-scale trial it was necessary for gathering treatment and baseline data (Valero-Mora et al., 2010).

Camera problems: Lighting differences among streets occurred to be a source of problems for recording the videos. Setting camera parameters to the most common levels of glare (in the cases where this was possible) produced videos that were non usable in narrow and dark streets. This problem affected all the cameras but was even more important for the camera that displayed a general view of the scenario, as it did not adjust automatically for brightness. Additionally, cameras that adjusted automatically for brightness presented their own problems. As an example, if the driver set the sunshade, the face camera received much reflection and the recording turned completely dark (a black plastic bag covering the sunshade fixed this problem) (Valero-Mora et al., 2010).

Driver behaviour during observation: The drivers stopped the car for entering a new destination in the navigator or for making or receiving a call. We had some doubts that their behaviour was completely naturalistic. Due to the short trips (about 2hours per day) with the unfamiliar car and its equipment drivers might not have forgotten the about observation and not behaved normal (Valero-Mora et al., 2010).

Data analysis: We had about 40 hours of driving data to analyze. The aim was to find critical incidents while driving (e.g. near-crashes, crash-relevant conflicts, illegal manoeuvres). The first approach was to first attempt consisted in setting thresholds for the numerical parameters that would be indicative of interesting events. The idea is to define trigger variables that point to interesting events, such as a large deceleration that might indicate a near collision. The problem is to define appropriate thresholds so that too many false positives are not produced but the important events are still detected (Groenewoud et al., 2010). In our case, we started with values that looked reasonable according to our experiences with the car as thresholds should be defined for each individual car or model of car. For example, as the equipment of the car ARGOS makes the car very heavy, the drivers needed to put more brake pressure than would be required in a standard car of the same size (Valero-Mora et al., 2010).

Unfortunately, the results were rather disappointing. First of all, we identified many incidents

which in their majority emerged as false alarms. As a typical example, sudden speed changes are common when stopping at a traffic light and driving in urban roads produced many of these. Secondly, there were real incidents that did not

show up when analyzing the numerical data. Moreover, many of the parameters are related with actions of the driver on the instruments of the car (brake, throttle, etc.) and in many incidents there is no such action, either because the driver lacks awareness of the critical situation or because there is no need of performing any extreme action (for example, braking softly may be sufficient) (Valero-Mora et al., 2010).

An analysis afterwards showed that none of the incidents had extraordinarily different values. A huge amount of false alarms would have to be analysed in order to find these incidents in the data whereas some wouldn't have been detected at all by just putting thresholds. The main issues are the following:

- *High number of false alarms:*
  - Urban traffic is full of conspicuous driving values.
  - Urban traffic leads constantly to sharp speed changes (e.g. at traffic lights).
  - Braking pressure seems to be in general high in Spanish traffic (e.g. due to much sudden braking).
  - Close distances between road users are usual in urban traffic (e.g. at traffic lights).
- *Missing incidents in driving data:*
  - Driving in lower urban speed does not always provoke extraordinary values for e.g. braking.
  - The other participant(s) of a conflictive incident perform(s) (predominantly) an evasive manoeuvre.

As a consequence the fourty hours of video material have been analyzed manually in order to identify the incidents. Due to technical issues and the in-depth analysis of certain moments of driving this analysis required about 2-2,5 hours for 1 hour of video-material. However, even a more efficient video analysis is not feasable in long-scale field trials.

CONCLUSION

The main goal of this small-scale field trial has been the learning about using a highly instrumented car for conducting a long-scale naturalistic driving study. While conducting the study, some technical processes were optimized, several unexpected issues were solved, methods for data reduction and analyze were tested and in consequence a number of lessons were learned that will be useful for large-scale studies.

In conclusion naturalistic driving methodology enables unobtrusive, realistic driving observation what is a great advantage compared to traditional methods. However, one important result is that the naturalistic character of studies with the ARGOS or similar cars requires some improvements for further studies. Thus, the driver should be less aware of all the instruments and systems around him in the car.

Data analysis is in general of special concern as a complete video analysis isn't feasible in large scale studies and filtering the numerical data requires improvement for finding critical moments in traffic. This issue is not limited to highly instrumented cars.

In sum, highly instrumented cars are with some improvements good instruments for naturalistic driving as they supply an extensive amount of information about the driver and his environment and in consequence enable an in-depth analysis of interesting moments in traffic.

AUTHOR'S NOTE

Acknowledgements: This work has been realized thanks to funding obtained by the European Commission FP-7 and the Directorate of Traffic (DGT) who made us available their instrumented vehicle ARGOS.

REFERENCES

Backer-Grøndahl, A., Phillips, R., Sagberg, F., Touliou, K., Gatscha, M. (2009). *Naturalistic driving observation: Topics and applications of previous and current naturalistic studies. PROLOGUE Deliverable D1.1.* TØI Institute of Transport Economics, Oslo, Norway.

Groenewoud, C., Schoen, E., Malone, K., Jonkers, E., Hoedemaeker, M., Hogema, J. (2010). *Methodological and organizational issues and requirements for ND studies.* Deliverable D2.2. PROLOGUE project.

Valero-Mora, P. M., Tontsch, A., Pareja-Montoro, I., Sánchez-García, M. (2010). *Using a highly instrumented car for naturalistic driving research: a small-scale study in Spain.* PROLOGUE Deliverable D3.5. INTRAS/University of Valencia, Valencia, Spain.

ÍNDICE DE AUTORES

Isasi, X.
Jara, P.
Lertxundi, A.
Lertxundi, N.
Livacic-Rojas, P.E.
López López, J. A.
López Pina, J. A.
López, E.
López, O.
López-Castedo, A.
Mahou, X.
Malapeira, J. M.
Manzano, L.
Marín Martínez, F.
Martínez, Z.
McArdle, J. J.
Molina, J. F.
Morales Sánchez, V.
Núñez Núñez, R. M.
Oliver, J. C.
Padilla, J. L.
Pallarés, J.
Páramo, M. F.
Pareja-Montoro, I.
Penelo, E.
Peña-Suárez, E.
Pérez Moreno, P. J.
Pérez-Moreno, E.
Peró-Cebollero, M.
Pirla, L.
Portell, M.
Prieto Marañón, P.
Privado, J.
Raedeke, T.
Raich, R. M.
Rangel Gascó, A.
Recarte, M. A.
Rial-Boubeta, A.
Rodríguez, M. S.
Romero Galisteo, R. P.
Rosa Alcázar, A. I.
Rosel, J.
Sánchez Guerrero, E.
Sánchez Meca, J.
Sánchez, I.
Sánchez-García, M.
Sánchez-Martín, M.
Sanduvete-Chaves, S.
Sarapura, R.
Seoane, G.
Serrano Gómez, V.
Sevillano, J. M.
Sireci, S. G.
Sueiro, M.
Suero, M.
Tinajero, C.
Tontsch, A.
Torregrosa, M.
Tuero-Herrero, E.
Valero-Mora, P.M.
Vallejo, G.
Van Den Noortgate, W.
Vegas, O.
Vergara, A.I.
Vergara, S.
Viader, M.
Viladrich, C.
Vozmediano, L.

El Congreso de Metodología de las Ciencias Sociales y de la Salud se organiza bianualmente con el auspicio de la Asociación Española de las Ciencias del Comportamiento (AEMCCO). Su objetivo consiste en fomentar el intercambio de conocimientos entre investigadores, docentes y profesionales pertenecientes a distintas disciplinas pero con un interés común en los avances metodológicos que permitan trabajar con rigor científico. En este libro de actas se incluyen algunos de los trabajos presentados en la XII edición del Congreso, que se celebró en la Facultad de Psicología de la Universidad del País Vasco / Euskal Herriko Unibertsitatea, entre los días 19 y 22 de Julio de 2011.

PATROCINADORES

Psikologia Fakultatea
Facultad de Psicología

Donostiako Udala
Ayuntamiento de
San Sebastián

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

EUSKO JAURLARITZA
GOBIERNO VASCO

aemcco

gizarte kutxa