

Evaluación de las diferencias individuales en falseamiento

Pere Joan Ferrando Piera y Cristina Anguiano Carrasco

Universidad Rovira i Virgili

En los últimos años, las puntuaciones de cambio obtenidas entre condiciones neutrales y condiciones de presión se han convertido en una de las principales alternativas para medir el falseamiento. Una importante cuestión aún pendiente relacionada con dichas puntuaciones es la relevancia de las diferencias individuales cuando las condiciones de presión son similares. El presente estudio propone un procedimiento, relativamente simple, y basado en la teoría clásica del test, para evaluar de forma rigurosa esta cuestión. El procedimiento, del que se derivan tres índices, se basa en un diseño pre-test post-test con grupo control y modela la magnitud de cambio como un parámetro individual. La propuesta se aplica en un estudio empírico en personalidad que proporciona resultados iniciales de interés.

Assessment of individual differences in faking. In recent years, change scores obtained under neutral conditions and under faking-inducing conditions have become one of the main alternatives for operationalizing faking. A pending issue regarding these measures is the relevance of individual differences under similar conditions of pressure. This study proposes a simple approach based on the classic test theory that allows the issue to be rigorously assessed. The approach, from which three indices are derived, is based on a pre-test post-test design with a control group, and models the amount of change as an individual parameter. The proposal is applied to an empirical study in personality, and some interesting initial results are obtained.

El falseamiento positivo (faking good) de un test de personalidad puede definirse como una distorsión deliberada de las respuestas, especialmente bajo condiciones de presión, a fin de crear una imagen más positiva o favorable (e.g., Griffith y Peterson, 2008; McFarland y Ryan, 2006). Se trata de un tópico muy relevante y que ha sido objeto de una gran cantidad de investigación. Además, el interés por él ha resurgido de forma espectacular en las dos últimas décadas de forma consecuente al renovado interés en el uso de medidas de personalidad en selección (e.g., Rothstein y Goffin, 2006).

Una diferencia interesante entre la investigación tradicional y la más reciente está en la definición operativa que priorizan. Existen dos enfoques principales para medir el falseamiento (McFarland, y Ryan, 2000, 2006). El primero, más indirecto, consiste en utilizar las puntuaciones en una escala de deseabilidad social (e.g., Ferrando y Chico, 2000; Salgado, 2005). El segundo, más directo, consiste en utilizar puntuaciones de cambio en una medida administrada en condiciones neutrales y en condiciones de presión para falsear. Ambos se han venido utilizando desde los inicios de la investigación sobre el tema. Sin embargo, la investigación tradicional ha priorizado el primer enfoque (Furnham, 1989), mientras que la moderna se ha basado cada vez más en las puntuaciones de cambio (Griffith y Peterson, 2008; McFarland y Ryan, 2006). A

pesar de este creciente interés, existen aún cuestiones pendientes acerca de los resultados obtenidos con las puntuaciones de cambio y su interpretación (Burns y Christiansen, 2006), una situación que puede ser debida en parte a problemas metodológicos: el análisis de las puntuaciones de cambio siempre ha sido un tema espinoso en Psicometría (e.g., Muñiz, 1992).

De acuerdo con las revisiones actuales (Burns y Christiansen, 2006; Mesmer-Magnus y Viswesvaran, 2006; Rothstein y Goffin, 2006), una de las cuestiones pendientes más básica es el efecto de las diferencias individuales en la magnitud del cambio detectado bajo condiciones similares de presión. Más en detalle, interesa conocer si parecidas condiciones instruccionales o motivacionales producen un cambio aproximadamente constante en todos los sujetos, o, por el contrario, la magnitud del cambio es en gran parte idiosincrásica. Esta cuestión, que es la que abordaremos aquí, tiene importancia tanto teórica como práctica. Desde el punto de vista teórico, por ejemplo, interesa conocer si tal efecto depende de la dimensión que se mide o del tipo de medida utilizado. La relevancia práctica, por otra parte, es inmediata (Furnham 1986; McFarland y Ryan 2000). Si, bajo las mismas condiciones, todos los individuos incrementan o disminuyen su puntuación en la misma cuantía, el rango de puntuaciones se mantendrá inalterado. Por tanto, las decisiones en un proceso de selección basado en este rango serán esencialmente correctas y también lo serán los coeficientes de validez con respecto a otras variables. Sin embargo, si unos individuos cambian sustancialmente más que otros, las decisiones pueden ser erróneas y los coeficientes de validez sesgados.

La revisión de los estudios existentes sugiere que los procedimientos de investigación utilizados son mejorables. La mayor parte de ellos se han basado en la correlación entre las puntuaciones

Fecha recepción: 16-12-10 • Fecha aceptación: 16-3-11

Correspondencia: Pere Joan Ferrando Piera

Facultad de Psicología

Universidad Rovira i Virgili

43007 Tarragona (Spain)

e-mail: perejoan.ferrando@urv.net

obtenidas en condiciones neutrales y en condiciones de presión. Como se ha discutido arriba, un cambio constante no alterará dicha correlación, mientras que los efectos diferenciales la atenuarán, y la atenuación será tanto más acusada cuanto mayor el efecto de las diferencias individuales. En general, los estudios han obtenido correlaciones claramente inferiores a las fiabilidades estimadas en cada ocasión separadamente (Braun, 1962; Mueller-Hanson, Heggstad y Thornton, 2006; Loo y Wudel, 1979). Estos resultados se han interpretado como evidencia de una variabilidad sustancial en los niveles de cambio debida a las diferencias individuales. En conjunto, tal como critican Gordon y Gross (1978), el enfoque correlacional es útil pero indirecto.

Gordon y Gross (1978) y Lautenschlager (1986) consideraron que el mejor índice para evaluar el efecto de las diferencias individuales era la varianza de las puntuaciones de cambio. Sin embargo, este índice parece haber sido muy poco utilizado en aplicaciones (Loo y Wudel, 1979). Además, tal como sus propios proponentes reconocen (Gordon y Gross, 1978), aunque la varianza es una medida más directa que la correlación, tiene problemas importantes. En efecto, es de esperar que dicha varianza no solo refleje la variabilidad debida a las diferencias individuales en el falseamiento, sino también la debida a efectos de memoria o retest, fluctuación temporal y error de medida.

El presente estudio propone un procedimiento general y una serie de índices para evaluar de forma rigurosa el efecto de las diferencias individuales en la magnitud del cambio atribuible al falseamiento. El procedimiento utiliza como medida las puntuaciones de cambio y se basa en un modelo simple enmarcado en la teoría clásica del test (e.g., Muñiz, 1992). A este respecto cabe considerar que existen metodologías teóricamente superiores desde las que se podría abordar el estudio, tales como los modelos longitudinales de ecuaciones estructurales (e.g., Ferrando y Anguano-Carrasco, 2011). La revisión de la literatura, sin embargo, indica que las investigaciones aplicadas se basan en su mayor parte en las puntuaciones simples de cambio que consideramos aquí, mientras que los procedimientos más complejos solo aparecen esporádicamente en las revistas especializadas en metodología. En suma, no se discute la potencial superioridad de otros posibles procedimientos más complejos. La intención es proponer una metodología básica, útil y relativamente fácil de utilizar en investigación aplicada, y que pueda utilizarse con el tipo de puntuaciones empleado habitualmente y con muestras no excesivamente grandes.

Además de la propuesta metodológica, el presente estudio pretende aportar una contribución empírica. Los procedimientos que se proponen se utilizan en un estudio en personalidad que, como se discute más adelante, es más que una simple ilustración del método y que puede aportar resultados iniciales de interés.

Diseño, modelo y propuesta

Con diferencia, el diseño más utilizado en la investigación aplicada para evaluar el cambio debido al falseamiento es el pre-test-post-test de un solo grupo, en el que se administra el test la primera vez bajo condiciones neutrales y la segunda vez bajo presión o instrucción para el falseamiento. Como es bien conocido, este diseño es muy débil, por ser susceptible a efectos de retest, historia y maduración (Shadish, Cook y Campbell, 2002). Mesmer-Magnus y Viswesvaran (2006) han recomendado vivamente a los investigadores que utilicen un diseño pre-test-post-test con un grupo de control al que se administraría la medida en condiciones neutrales

en las dos ocasiones. Este segundo diseño se considera ya generalmente interpretable (Shadish, Cook y Campbell, 2002) y es el que consideraremos aquí. Más en detalle, las condiciones que asumiremos son las siguientes. En primer lugar, que los grupos experimental y control son lo bastante equivalentes como para permitir comparaciones. En segundo lugar, que el intervalo de retest es el mismo en los dos grupos y es lo bastante largo como para hacer negligibles los efectos de memoria. Por último, por claridad en la exposición, consideraremos que el test se puntúa en la dirección más positiva, adaptativa o socialmente deseable del rasgo, de forma que los cambios esperados son siempre hacia arriba.

Conceptualmente, la propuesta se basa en el modelo de cambio de Zickar y Robie (1999), el cual plantea que bajo condiciones de presión para el falseamiento se produce una elevación temporal en el nivel o puntuación verdadera del sujeto que lleva a puntuaciones más positivas en los ítems. La magnitud de esta elevación temporal se modela como un parámetro individual, y la variabilidad en los valores de dicho parámetro es el objeto de estudio del procedimiento.

Sea $X_{ik}^{(g)}$ la puntuación observada en el individuo i perteneciente al grupo $g = c(\text{control})$ o $e(\text{experimental})$ en el momento k (1 pre-test o 2 post-test). El modelo de respuesta en el pre-test es el modelo clásico de puntuación verdadera (e.g., Muñiz, 1992):

$$X_{i1}^{(c)} = T_{i1}^{(c)} + \varepsilon_{i1}^{(c)}; X_{i1}^{(e)} = T_{i1}^{(e)} + \varepsilon_{i1}^{(e)} \quad (1)$$

Para el grupo de control, el modelo en el post-test es:

$$X_{i2}^{(c)} = T_{i2}^{(c)} + \varepsilon_{i2}^{(c)} = (\beta T_{i1}^{(c)} + \omega^{(c)}_i) + \varepsilon_{i2}^{(c)} \quad (2)$$

que es un modelo longitudinal estándar de cambio (e.g., Kenny y Campbell, 1989). La puntuación verdadera en el post-test se descompone en dos componentes incorrelados: la parte que puede predicirse desde los niveles verdaderos iniciales y un componente aleatorio ω que modela las influencias tales como fluctuación o erosión temporal, que pueden haber afectado al rasgo durante el período de retest (veáanse e.g. Kenny y Campbell, 1989; o Tucker, Damarin y Messick, 1966). La pendiente de regresión de T_2 sobre T_1 se supone invariante en ambos grupos, y, por tanto, se representa sin el superíndice. Este supuesto es razonable cuando ninguno de los grupos es seleccionado en base a sus puntuaciones iniciales (Shadish, Cook y Campbell, 2002).

El modelo correspondiente para el grupo experimental es:

$$X_{i2}^{(e)} = T_{i2}^{(e)} + \varepsilon_{i2}^{(e)} = (\beta(T_{i1}^{(e)} + \delta_i) + \omega^{(e)}_i) + \varepsilon_{i2}^{(e)} \quad (3)$$

Donde el término δ_i modela la magnitud del cambio debido al falseamiento en el individuo i con respecto a su nivel inicial.

La puntuación de cambio se define ahora como la diferencia entre la puntuación observada en el post-test y la observada en el pre-test: $D_i = X_{i2} - X_{i1}$, y el procedimiento que se propone se basa en las varianzas esperadas de dichas puntuaciones en cada uno de los grupos. Para desarrollarlas, se usan una serie de restricciones razonables. En primer lugar, la varianza de los errores de medida se considera dependiente del test, y, por tanto, se considera la misma para ambos grupos y condiciones. En segundo lugar, la varianza de las puntuaciones verdaderas en el pre-test se considera la misma

en ambos grupos (ya que son equivalentes y comparables). Por último, se supone que la varianza del término de error ω es la misma en ambos grupos, ya que el período de retest es el mismo y no se esperan diferencias sistemáticas en fluctuación entre grupos. Si se omiten los subíndices y superíndices correspondientes a estas restricciones, las varianzas quedan como:

$$\sigma^2(D^{(c)}) = (\beta - 1)^2 \sigma^2(T_1) + \sigma^2(\omega) + 2\sigma^2(\varepsilon)$$

$$\sigma^2(D^{(e)}) = (\beta - 1)^2 \sigma^2(T_1) + \beta^2 \sigma^2(\delta) + \sigma^2(\omega) + 2\sigma^2(\varepsilon) \quad (4)$$

La segunda de las varianzas en (4) corresponde al indicador propuesto por Gordon y Gross (1978) y Lautenschlager (1986). Sin embargo, como se ha discutido arriba, el componente de varianza que interesa evaluar es $\sigma^2(\delta)$: la variabilidad en los niveles de cambio atribuible al falseamiento y libre de los efectos de error y fluctuación. Comparando los desarrollos en ambos grupos, dicho componente viene dado según el modelo por:

$$\sigma^2(\delta) = \frac{\sigma^2(D^{(e)}) - \sigma^2(D^{(c)})}{\beta^2} \quad (5)$$

De los resultados descritos hasta ahora se derivan los tres índices que se quieren proponer. El primero de ellos es el estimador de $\sigma^2(\delta)$ en (5) obtenido en las dos muestras o grupos bajo estudio. Los términos de varianza del numerador de (5) se estiman mediante las varianzas de las puntuaciones de cambio observadas en ambos grupos. La pendiente de regresión de T_2 sobre T_1 se estima en el grupo de control mediante la pendiente habitual de regresión corregida por atenuación (e.g., Tucker, Damarin y Messick, 1966):

$$\hat{\sigma}^2(\delta) = \frac{s^2(D^{(e)}) - s^2(D^{(c)})}{\hat{\beta}^2}; \quad \hat{\beta} = \frac{r_{xx_1}^{(c)} s^{(c)}(X_2)}{s^{(c)}(X_1)} \quad (6)$$

donde r_{xx_1} es la fiabilidad estimada de las puntuaciones en el pre-test.

Las principales limitaciones del primer índice propuesto son las de toda varianza en general: (a) su valor depende de la métrica del test, y (b) no tiene límite superior, lo que hace que sea difícil de interpretar. Gordon y Gross (1978) propusieron establecer valores de referencia a partir de los resultados empíricos, pero esto requiere una cierta acumulación de estudios. En principio, hay dos importantes situaciones para las que el índice es potencialmente útil: para evaluar si difiere significativamente de cero (la hipótesis nula de cambio constante en todos los sujetos) y para comparar con el mismo test la variabilidad individual producida bajo diversas condiciones experimentales (e.g., niveles de presión o diferentes instrucciones).

Para paliar las limitaciones del índice principal se hacen dos propuestas adicionales que ayuden a su interpretación. En primer lugar, la raíz cuadrada positiva del índice (6) puede interpretarse como una desviación típica y está en las mismas unidades de medida que las puntuaciones de cambio. El cociente entre esta desviación y la media de las puntuaciones de cambio es un coeficiente de variación, no depende de las unidades de medida del test, y

refleja la magnitud de la variabilidad individual inducida por el falseamiento en relación a la tendencia media de cambio observada globalmente en el grupo.

La segunda propuesta adicional es calcular el cociente entre el componente de cambio debido al falseamiento y toda la variabilidad verdadera debida a las diferencias individuales. El correspondiente tercer índice propuesto es:

$$RV(\delta) = \frac{\hat{\sigma}^2(\delta)}{\hat{\sigma}^2(T_1) + \hat{\sigma}^2(\omega) + \hat{\sigma}^2(\delta)} \quad (7)$$

Se trata de un índice de varianza relativa que indica qué proporción de la variabilidad debida a las diferencias individuales (libre de error de medida) es atribuible al falseamiento. Como variabilidad total verdadera tomamos: (a) la debida a las diferencias individuales en el pre-test; (b) la debida a la fluctuación temporal; y (c) la debida exclusivamente al falseamiento. El término (a) de varianza en el denominador puede obtenerse multiplicando la varianza observada en el pretest por la fiabilidad. Una vez obtenido, el segundo término (b) puede calcularse a partir de cualquiera de las dos ecuaciones en (4).

Para contrastar la hipótesis de cambio constante o establecer comparaciones en base a los índices (6) y (7) es necesario determinar un margen de error probable basado en su distribución muestral. La determinación analítica de tal distribución es extremadamente compleja, ya que en ambos (6) y (7) tanto los elementos del numerador como los del denominador son estadísticos sujetos a fluctuación muestral. La solución que proponemos es determinar una distribución muestral empírica mediante réplicas autodocimantes (Bootstrap) y obtener el intervalo de confianza al nivel deseado mediante los percentiles de dicha distribución. En el estudio que sigue se utilizaron 500 replicaciones en todos los casos.

Estudio empírico

Existen dos tipos principales de estudios empíricos sobre falseamiento: estudios en situaciones reales de selección y estudios en los que la presión a falsear se provoca mediante instrucciones. Los dos tipos son complementarios, proporcionan información distinta y permiten un distinto grado de control (Mesmer-Magnus y Viswesvaran, 2006). Así, los estudios de falseamiento instruccional pueden considerarse evaluaciones de rendimiento máximo, midiendo los límites del falseamiento (Salgado, 2005). Permiten también un mayor control experimental, ya que todos los sujetos reciben exactamente las mismas instrucciones. Esta segunda característica es muy relevante en el presente caso, ya que el papel de las diferencias individuales se puede evaluar de forma más clara cuando las condiciones de presión son las mismas para todos. El presente estudio es de falseamiento instruccional.

El estudio se basa en dos dimensiones de personalidad de tipo cuasi-patológico: neuroticismo y psicoticismo. Existe una gran cantidad de investigación acerca del falseamiento en estas dimensiones y, además, los resultados a nivel grupal son muy consistentes (Furnham, 1986). En situaciones de presión se producen cambios sustanciales en la dirección de mayor estabilidad emocional (dimensión de Neuroticismo), y mayor apertura, empatía y sensibilidad (Psicoticismo). Sin embargo, el papel de las diferencias individuales en estas dos tendencias generales no parece haber sido explorado, y éste será aquí el principal objetivo.

Método

Participantes y procedimiento

El grupo total estaba formado por 475 estudiantes de las facultades de Educación y Psicología de una universidad española, que fueron asignados a un grupo de control ($N=208$) y a un grupo experimental ($N=267$). La media de edad (21 años) y la distribución por sexos (80% mujeres) era la misma en ambos grupos. Las medidas se administraron en papel y lápiz, de forma voluntaria y anónima, en grupos de clase de 25 a 60 alumnos y siempre por el mismo investigador. La asignación de los participantes a los dos grupos fue cuasi-aleatoria y por conglomerados (Shadish, Cook y Campbell, 2002): los grupos clase fueron asignados al azar al grupo control o al experimental.

Las instrucciones en el pre-test eran las del manual del cuestionario, que recomiendan, entre otras cosas, responder honestamente. En el post-test, los participantes del grupo de control volvieron a responder bajo las mismas instrucciones. Los del grupo experimental recibieron las instrucciones especificadas en Eysenck, Eysenck y Shaw (1974). Se les pedía que se imaginasen como solicitantes a un puesto de trabajo en el que estaban realmente interesados y que trataran de responder, no de forma honesta, sino intentando dar la mejor impresión posible, respondiendo de acuerdo con lo que creían que al seleccionador le gustaría. Con respecto al segundo requisito previo discutido arriba, el intervalo de retest fue el mismo en todos los casos: 6 semanas. En la experiencia de los autores este intervalo es suficiente para evitar efectos de memoria en el tipo de medidas aquí utilizado.

Medidas

Se utilizaron las escalas de Neuroticismo (24 ítems) y Psicoticismo (32 ítems) de la adaptación española del EPQ-R llevada a cabo por Aguilar, Tous y Andrés (1990). Tal como se especificaba en el tercer requisito previo, las escalas se puntuaron en la dirección más socialmente deseable del rasgo.

Resultados

La tabla 1 muestra los descriptivos de ambos grupos en las dos mediciones. Para ambas medidas, los descriptivos en el pre-test (medias, desviaciones y fiabilidades) son muy similares en ambos grupos. Adicionalmente, la prueba de comparación de medias entre ambos grupos en T1 no dio resultados significativos en ninguna de las dos escalas ($N: t=0.82; p=0.41; y P: t=0.61; p=0.54$). Así pues, los resultados apoyan el primer requisito previo de comparabilidad inicial entre grupos. Con respecto a los cambios pre-post test, los resultados en el grupo de control sugieren que no hay ninguna tendencia de cambio significativa en ninguna de las dos medidas. Por el contrario, en ambas se aprecia un cambio notable en la dirección esperada en el grupo experimental. La diferencia media es de aproximadamente una unidad de desviación en el caso de Neuroticismo y de 0,8 unidades en el caso de Psicoticismo.

La tabla 2 muestra los resultados principales en relación a la presente propuesta. Con respecto al primer índice, el estimador de varianza de cambio debida al falseamiento, se observa que, en ambas escalas, el valor estimado difiere significativamente de cero de forma muy clara (el valor cero queda fuera del intervalo de confianza). Además las estimaciones puntuales son muy similares

en ambas medidas y los intervalos se solapan casi totalmente. La conclusión sería pues que las diferencias individuales tienen una influencia significativa en la magnitud del cambio debido al falseamiento y que dicha influencia, en términos absolutos, es la misma en ambas medidas.

Los indicadores relativos sugieren, en cambio, diferencias consistentes entre las dos medidas. La inspección de los coeficientes de variación obtenidos en relación a las medias de las diferencias al final de la tabla 1 sugiere que la variabilidad relativa es mayor en la medida de Psicoticismo. En otras palabras, en relación a la tendencia media de cambio observada en todo el grupo, las diferencias individuales tienen un peso bastante mayor en el caso del Psicoticismo que en el del Neuroticismo.

El tercer índice, la medida de razón varianza relativa (7), por último, va en la misma dirección. La estimación puntual sugiere que, en el caso del Neuroticismo, un 34% de la varianza verdadera de las diferencias individuales se debe a diferencias en los niveles de cambio individuales cuando se responde bajo las mismas instrucciones. En el caso del Psicoticismo, esta proporción se eleva al 55%. Como se discute más abajo, en ambos casos el efecto parece notable. Además, parece bastante claro que el efecto es sustancialmente mayor en el caso del Psicoticismo, ya que aquí los intervalos de confianza no se solapan.

Discusión y conclusiones

El primer objetivo de este trabajo era proponer una metodología a la vez simple y rigurosa para evaluar el papel de las diferencias individuales en la magnitud del cambio inducido por falseamiento. Consideramos que este objetivo se ha cumplido, y que los índices derivados de la propuesta pueden ser de utilidad en la investigación aplicada. Ahora bien, para que los resultados obtenidos con el método sean claros e interpretables deben cumplirse dos tipos

Tabla 1
Estadísticos descriptivos

		Neuroticismo		Psicoticismo	
		Control	Experimental	Control	Experimental
Pre-test (n= 475)	Media	12,64	13,01	24,18	24,44
	D. típica	5,11	4,76	3,41	3,40
	Fiabilidad	0,85	0,82	0,67	0,65
Post-test (n= 475)	Media	13,01	17,27	24,70	27,24
	D. típica	5,63	3,81	3,44	3,50
P. cambio	Media	0,37 (n.s.)	4,26	0,52 (n.s.)	2,80
	D. típica	3,55	4,89	2,4	4,33

Nota: n.s.: no significativo

Tabla 2
Evaluación del efecto de las diferencias individuales en el cambio debido al falseamiento

Medida	$\delta^2(\delta)$ y 90% I.C.	$\delta(\delta)$	C.deV.	RV(δ) y 90% I.C.
Neuroticismo	10,89 (7,52 ; 14,78)	3,30	0,77	0,34 (0,26 ; 0,41)
Psicoticismo	10,13(7,49 ; 13,58)	3,18	1,13	0,55 (0,48 ; 0,62)

generales de condiciones. Este cumplimiento es particularmente relevante en este caso ya que la propuesta se basa en la teoría clásica del test, y, por tanto, algunos de sus supuestos básicos no son empíricamente falsificables.

En primer lugar, la presente propuesta es simple principalmente por utilizar como medidas básicas las puntuaciones del test. Por tanto, su uso solo tendrá sentido si dichas puntuaciones son, de entrada, interpretables y cumplen una serie de condiciones psicométricas básicas, principalmente unidimensionalidad y precisión en la medida (e.g., Muñiz, 1992). En efecto, la ecuación (1) implica unidimensionalidad y esta condición se puede evaluar mediante procedimientos de análisis factorial. También puede ser importante evaluar la consistencia de las respuestas de los sujetos en el pre-test (véase Ferrando y Morales, 2010). Por último, las puntuaciones del test son necesariamente limitadas y, por tanto, susceptibles a efectos techo si bajo las condiciones experimentales una elevada proporción de sujetos puntúa en el límite superior. En este tipo de situaciones sería posible introducir correcciones basadas en los modelos de variables censuradas (véase Ferrando y Lorenzo-Seva, 1999). Sin embargo, la experiencia de los autores con diferentes conjuntos de datos sugiere que los cambios que las correcciones producen en los índices aquí propuestos son generalmente mínimos.

El segundo tipo de requisitos básico refiere a la comparabilidad de los grupos control y experimental. La simplicidad de los índices se deriva también en gran parte de las restricciones de equivalencia entre grupos: la misma varianza verdadera en el pretest, la misma varianza debida a fluctuación en ambos y pendiente de regresión desatenuada invariante. La primera restricción puede evaluarse mediante análisis factorial simultáneo en los dos grupos. Sin embargo, las otras dos no son, en principio, verificables y la mejor opción es reforzar su plausibilidad mediante el control experimental. Más en detalle, asegurar en lo posible la equivalencia mediante asignación aleatoria o cuasi-aleatoria (por ejemplo de grupos naturales de sujetos) de los participantes a los grupos (pendiente invariante) y procurar que el intervalo de retest sea el mismo (idéntica fluctuación). No se espera que el método funcione si los grupos no son equivalentes, si el período de retest es distinto en ambos, o, como se ha dicho antes, si los grupos se seleccionan en base a sus puntuaciones iniciales.

Pasamos a discutir los resultados del estudio empírico que era el segundo objetivo. Furnham (1986) conjeturó que el papel de las diferencias individuales ha de ser más relevante en aquellas

dimensiones donde la dirección del falseamiento es menos clara. De ser así, los presentes resultados son notables, ya que se han escogido dos medidas cuyo polo inferior es claramente desadaptativo y roza la patología. En otras palabras, la dirección hacia donde falsear es aquí, a priori, muy clara y la tendencia detectada a nivel de grupo bastante fuerte. Sin embargo, aun en estas condiciones, la variabilidad estimada debida a las diferencias individuales ha sido considerable. Con las necesarias precauciones, este resultado sugiere que el objeto de estudio de la propuesta no es trivial y merece investigación futura. También parece un resultado de interés la mayor variabilidad relativa estimada en el caso del Psicoticismo. La escala de Psicoticismo ha sido criticada por su heterogeneidad de contenido y baja precisión (Chico y Ferrando, 1995). Es posible que estos factores hagan menos clara la dirección del falseamiento en relación a la más consistente medida de Neuroticismo, y, en consecuencia, las diferencias individuales adquieran mayor relevancia, pero esto es tan solo una conjetura. En términos mucho más generales, creemos que las futuras líneas de investigación en relación a estos resultados iniciales podrán ir en dos direcciones. En primer lugar, investigar el papel de las diferencias individuales a través de diferentes dimensiones y formas de medirla (incluyendo tipos de ítems, formato de respuesta, etc.). En segundo lugar, investigar dicho papel tanto en condiciones de selección como instruccionales, variando las condiciones experimentales. Por ejemplo, presión moderada vs presión alta en el primer caso, o instrucciones genéricas (las utilizadas en nuestro estudio) versus condiciones específicas en el segundo.

Finalmente, la experiencia sugiere que las propuestas del tipo presentado solo se acaban utilizando ampliamente si son sencillas de calcular o están implementadas en programas accesibles y, a ser posible, de libre distribución. Con respecto al primer punto, los tres índices propuestos pueden obtenerse fácilmente mediante una hoja de cálculo o un programa estadístico genérico. Los intervalos mediante replicación, sin embargo, son más complejos. El primer autor ha desarrollado programas en el entorno MATLAB que implementan la propuesta y que pone a disposición de los lectores que lo soliciten.

Agradecimientos

Esta investigación ha sido parcialmente subvencionada por el Ministerio de Ciencia y Educación (PSI2008-00236/PSIC).

Referencias

- Aguilar, A., Tous, J.M., y Andrés, A. (1990). Adaptación y estudio psicométrico del EPQ-R. *Anuario de Psicología*, 46, 101-118.
- Braun, J.R. (1962). Differential susceptibility to faking of various Ghiselli Self-description inventory scales. *Psychological Reports*, 10, 639-641.
- Burns, G.N., y Christiansen, N.D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. En R.L. Griffith y M.H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 113-148). Greenwich, CT: Information Age Publishing.
- Chico, E., y Ferrando, P.J. (1995). A psychometric evaluation of the revised P scale in delinquent and non-delinquent Spanish samples. *Personality and Individual Differences*, 18, 331-337.
- Eysenck, S.B.G., Eysenck, H.J., y Shaw, L. (1974). The modification of personality and Lie scores by special 'honestly' instructions. *British Journal of Social and Clinical Psychology*, 13, 41-50.
- Ferrando, P.J., y Anguiano-Carrasco, C. (2011). A structural equation model at the individual and group level for assessing faking-related change. *Structural Equation Modeling*, 18, 1-19.
- Ferrando, P.J., y Chico, E. (2000). Adaptación y análisis psicométrico de la escala de deseabilidad social de Marlowe y Crowne. *Psicothema*, 12, 383-389.
- Ferrando, P.J., y Lorenzo-Seva, U. (1999). Implementing a test of underlying normality for censored variables. *Multivariate Behavioral Research*, 34, 421-439.
- Ferrando, P.J., y Morales, F. (2010). Procedimientos gráficos para evaluar la consistencia intra-individual en el análisis factorial de ítems. *Psicothema*, 22, 348-354.
- Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7, 385-400.

- Gordon, M.E., y Gross, R.H. (1978). A critique of methods for operationalizing the concept of fakeability. *Educational and Psychological Measurement*, 38, 771-782.
- Griffith, R.L., y Peterson, M.H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology*, 1, 308-311.
- Kenny, D.A., y Campbell, D.T. (1989). On the measurement of stability in over-time data. *Journal of Personality*, 57, 445-481.
- Lautenschlager, G.J. (1986). Within-subject measures for the assessment of individual differences in faking. *Educational and Psychological Measurement*, 46, 309-316.
- Loo, R., y Wudel, P. (1979). Estimates of fakeability on the Eysenck personality questionnaire. *Social Behavior and Personality: An International Journal*, 7, 157-160.
- McFarland, L.A., y Ryan, A.M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, 85, 812-821.
- McFarland, L.A., y Ryan, A.M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology*, 36, 979-1016.
- Mesmer-Magnus, J., y Viswesvaran, C. (2006). Assessing response distortion in personality tests: A review of research designs and analytic strategies. En R.L. Griffith y M.H. Peterson (Eds.), *A closer examination of applicant faking behaviour* (pp. 85-113). Greenwich, CT: Information Age Publishing.
- Mueller-Hanson, R.A., Heggstad, E.D., y Thorton, G.C. (2006). Individual differences in impression management: An exploration of the psychological process underlying faking. *Psychology Science*, 48, 288-312.
- Muñiz, J. (1992). *Teoría clásica de los tests*. Madrid: Pirámide.
- Rothstein, M.G., y Goffin, R.D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, 16, 155-180.
- Salgado, J.F. (2005). Personalidad y deseabilidad social en contextos organizacionales: implicaciones para la práctica de la psicología del trabajo y las organizaciones. *Papeles del Psicólogo*, 92, 1-8.
- Shadish, W.R., Cook, T.D., y Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Tucker, L.R., Damarin, F., y Messick, S. (1966). A base-free measure of change. *Psychometrika*, 31, 457-473.
- Zickar, M.J., y Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, 84, 551-563.