Papeles del Psicólogo / Psychologist Papers, 2022 Vol. 43(1), pp. 29-35
https://doi.org/10.23923/pap.psicol.2982
http://www.papelesdelpsicologo.es
http://www.psychologistpapers.com

Artículo en prensa

CONSTRUYENDO TEST ADAPTATIVOS DE ELECCIÓN FORZOSA "ON THE FLY" PARA LA MEDICIÓN DE LA PERSONALIDAD

BUILDING ADAPTIVE FORCED CHOICE TESTS "ON THE FLY" FOR PERSONALITY MEASUREMENT

Abad, Francisco J.[1]; Kreitchmann, Rodrigo S.[1]; Sorrel, Miguel A.[1]; Nájera, Pablo[1]; García-Garzón, Eduardo[2]; Garrido, Luis Eduardo[3] y Jiménez, Marcos[1]

[1]Universidad Autónoma de Madrid; [2]Universidad Camilo José Cela; [3]Pontificia Universidad Católica Madre y Maestra, República Dominicana

Los nuevos desarrollos metodológicos y tecnológicos de la última década permiten resolver, o al menos atenuar, los problemas psicométricos de los test de elección forzosa (EF) para la medición de la personalidad. En estas pruebas, a la persona evaluada se le muestran bloques de dos o más frases de parecida deseabilidad social, entre las que debe elegir aquella que le represente mejor. De esta manera, los test de EF buscan reducir los sesgos de respuesta en pruebas de autoinforme. No obstante, su uso no está exento de riesgos y complicaciones si no se elaboran adecuadamente. Afortunadamente, los nuevos modelos psicométricos permiten modelar las respuestas en este tipo de test, así como optimizar su construcción. Más aún, permiten la construcción de Test Adaptativos Informatizados de EF (TAI-EF) "on-the-fly", en los que cada bloque se construye en el mismo momento de aplicación, emparejando óptimamente las frases de un banco previamente calibrado.
***Palabras clave**: Personalidad, Elección forzosa, Test adaptativo informatizado, On-the-fly.*

The new methodological and technological developments of the last decade make it possible to resolve or, at least, attenuate the psychometric problems of forced-choice (FC) tests for the measurement of personality. In these tests, the person being tested is shown blocks of two or more sentences of similar social desirability, from which he or she must choose which one best represents him or her. Thus, FC tests aim to reduce response bias in self-report questionnaires. However, their use is not without risks and complications if they are not created properly. Fortunately, new psychometric models make it possible to model responses in this type of test and to optimize their construction. Moreover, they allow the construction of "on the fly" computerized adaptive FC tests (CAT-FC), in which each item is constructed on the spot, optimally matching sentences from a previously calibrated bank.
***Key words:** Personality, Forced-choice, Computerized adaptive tests, On-the-fly.*

Diversos estudios de meta-análisis avalan el papel predictivo de las variables de personalidad, medidas por autoinforme, en contextos organizacionales, educativos y de la salud (p.ej., Barrick & Mount, 1991; Judge et al., 2013; Otero et al., 2020). Sabemos, por ejemplo, que la Responsabilidad y la Estabilidad Emocional tienen una validez predictiva generalizada a través de distintas ocupaciones y criterios, mientras que otras dimensiones, como la Apertura a la experiencia, la Afabilidad o la Extraversión, también resultan relevantes en contextos particulares y en la predicción de criterios específicos. En el ámbito educativo, la Responsabilidad y la Estabilidad emocional juegan un papel importante en la predicción del rendimiento académico (Poropat, 2009; Richardson et al., 2012), mientras que la responsabilidad y la afabilidad predicen conductas indebidas (p.ej., copiar en exámenes; Cuadrado et al., 2021). A pesar de estos resultados, las dudas sobre la sensibilidad del autoinforme a los efectos de la deseabilidad social y del falseamiento (*faking*) han acompañado a estas pruebas desde su creación. En concreto, en contextos de selección se espera de los candidatos cierta distorsión de las respuestas para dar una imagen más positiva de sí mismos, bien por autoengaño, bien deliberadamente para ser elegidos. Estas distorsiones producen fuertes incrementos en las medias de las puntuaciones en la dirección percibida como deseable y reducen la fiabilidad y la variabilidad de las puntuaciones (Viswesvaran & Ones, 1999; Salgado, 2016). Estos efectos se encuentran tanto en estudios experimentales, cuando se comparan respuestas honestas y con deshonestidad inducida, como, aunque más atenuados, en contextos aplicados, cuando se comparan muestras de solicitantes a puestos laborales y de personas que ya ocupan puestos laborales (Salgado, 2005). Por tanto, sin menoscabo de que en contextos reales se encuentren resultados positivos sobre la validez predictiva de las puntuaciones de personalidad, la evidencia previa sugiere que el ordenamiento de los candidatos podría ser distinto según falseen sus respuestas o no. Salgado (2005) describe algunas estrategias para reducir los efectos del falseamiento, tales como informar a los evaluados de que existe la posibilidad de ser penalizados si distorsionan sus respuestas o utilizar baremos específicos, confeccionados con muestras de solicitantes.

Otra posibilidad es utilizar formatos de respuesta más robustos al falseamiento. En el ámbito de las pruebas de autoinforme, puede distinguirse entre el formato tradicional de escala Likert (i.e., ítems o enunciados para los que el evaluado tiene que indicar su grado de acuerdo) y el formato de elección forzosa (i.e., bloques de uno o

Recibido: 1 noviembre 2021 - Aceptado: 16 diciembre 2021
Correspondencia: Abad, Francisco J.
E-Mail: fjose.abad@uam.es.

.....

Este trabajo fue realizado con la ayuda financiera del Ministerio de Ciencia, Innovación y Universidades (Proyecto PSI2017-85022-P) y por la Cátedra de Modelos y Aplicaciones Psicométricos (Instituto de Ingeniería del Conocimiento y Universidad Autónoma de Madrid).

más enunciados, entre los que el evaluado debe hacer una elección -p.ej., indicar el que mejor le representa- o un ordenamiento -p.ej., ordenarlos parcial o totalmente en función del grado en que le describen). En la Tabla 1 se muestran ejemplos de ítems de elección forzosa con distintos formatos.

El formato Likert es susceptible no sólo a los efectos del falseamiento o la deseabilidad social, sino también a la presencia de otros sesgos de respuesta tales como los sesgos de aquiescencia, negatividad, de tendencia central o respuesta extrema, etc. La existencia de sesgos de respuesta puede distorsionar la estructura factorial de la escala y derivar en desajuste (p.ej., Abad et al., 2018), además de producir una sobrestimación de la fiabilidad y alterar las estimaciones de validez convergente. Por el contrario, estos sesgos no son aplicables al formato de elección forzosa. Especialmente, se espera que, si los bloques se forman con ítems igualados en deseabilidad social, la susceptibilidad al falseamiento se reduzca.

PROBLEMAS DE LAS PUNTUACIONES OBTENIDAS EN PRUEBAS DE ELECCIÓN FORZOSA

A pesar de lo anterior, el uso del formato de elección forzosa no ha estado libre de controversia. En primer lugar, se ha puesto en duda su mayor resistencia al falseamiento (p.ej., Heggestad et al., 2006). No obstante, los estudios de meta-análisis más recientes sugieren que el efecto del falseamiento es menor en pruebas de elección forzosa (Cao & Drasgow, 2019; Martínez & Salgado, 2021). En segundo lugar, los test de elección forzosa pueden derivar en puntuaciones con propiedades *ipsativas*, en las que la interpretación de una puntuación es relativa al resto de puntuaciones del mismo sujeto. Por ejemplo, una persona muy organizada y muy sociable puede coincidir en su respuesta con otra poco organizada y poco sociable, por considerarse ambas *más* organizadas que sociables. Si las puntuaciones son completamente ipsativas, la suma de las puntuaciones de cada sujeto dará lugar a un mismo valor constante y las interpretaciones *normativas* (p.ej., concluir que la primera persona es más organizada que la segunda) serían arriesgadas. En estos casos, la aplicación de las técnicas tradicionales de análisis psicométrico derivará en artefactos metodológicos (Hicks, 1970). Por ejemplo, la correlación esperada promedio entre las dimensiones tenderá a ser negativa. De la misma manera, las correlaciones entre las puntuaciones en esas dimensiones y cualquier criterio externo será cero. Estos resultados proceden de las covarianzas negativas que se producen al forzar a una persona a elegir un enunciado frente a otro. Por ejemplo, suponga el caso extremo en el que un test incluya 20 bloques de dos ítems, uno puntuando positivamente en Extroversión y otro puntuando positivamente en Responsabilidad. Si sumamos +1 en Extroversión por cada ítem de Extroversión elegido y +1 en Responsabilidad por cada ítem de Responsabilidad elegido, la correlación entre ambas escalas será -1 y la suma de las puntuaciones en ambas escalas será 20, independientemente de las elecciones de los evaluados.

La ipsatividad no es una cuestión de todo o nada, ni va asociada al formato en sí, sino que depende del diseño del test y de los bloques (p.ej., número de ítems por bloque, naturaleza unidimensional/multidimensional de los bloques, polaridad directa/inversa de los ítems que forman los bloques, número de dimensiones evaluadas, correlación entre las dimensiones medidas o modo de puntuación). En este sentido, las puntuaciones en una prueba de elección forzosa pueden ser (Hicks, 1970): (a) totalmente ipsativas; (b) cuasi-ipsativas o parcialmente ipsativas; y (c) normativas. Las puntuaciones normativas pueden obtenerse, por ejemplo, si los ítems de un mismo bloque pertenecen a la misma dimensión. Las puntuaciones pueden hacerse parcialmente ipsativas si, por ejemplo, los evaluados ordenan parcialmente -más que completamente- las alternativas, las escalas difieren en el número de ítems, o una de las dimensiones no se puntúa. Las escalas cuasi-ipsativas dan lugar a puntuaciones que no suman una constante para todos los individuos, pero pueden mantener cierta interdependencia; esto es, el problema se reduce, pero

TABLA 1
EJEMPLOS DE ÍTEM CON FORMATO DE ELECCIÓN FORZOSA

Tipo de formato (entre paréntesis, para cada frase, dimensión medida y polaridad)	Elección/ puntuación	Elección/ puntuación
PICK.PAIR. *Elige la frase que mejor te representa:*	A /	B/
A. Creo que los demás tienen buenas intenciones (Af+)	Af: +1	Af: +0
B. Hago listas de cosas a hacer (Re+)	Re: +0	Re: +1
PICK. *Elige la frase que mejor te representa:*	A/	B/
A. Me siento relajado la mayor parte del tiempo (Es+)	Es: +1	Es: +0
B. Creo que los demás tienen buenas intenciones (Af+)	Af: +0	Af: +1
C. Hago listas de cosas a hacer (Re+)	Re: +0	Re: +0
MOLE (ej., Heggestad et al., 2006). *Indica las frases que mejor (↑) y peor (↓) te representan:*	C↑; B↓ /	B↑; C↓/
A. Evito el material de lectura difícil (Ap-)	Ap: +0	Ap: +0
B. Solo me siento cómodo con amigos (Ex-)	Ex: -(-1)	Ex: -(+1)
C. Creo que los demás tienen buenas intenciones (Af+)	Af: (+1)	Af: (-1)
D. Hago listas de cosas a hacer (Re+)	Re: +0	Re: +0
RANK. Ordena las frases según el grado en que te representan, desde "más parecido a ti" hasta "menos parecido a ti":	A>B>C>D/	D>B>C>A/
A. Me siento relajado la mayor parte del tiempo (Es+)	Es: +4	Es: +1
B. Creo que los demás tienen buenas intenciones (Af+)	Af: +3	Af: +3
C. Hago listas de cosas a hacer (Re+)	Re: +2	Re: +2
D. Me gusta aprender cosas nuevas (Ap+)	Ap: +1	Ap: +4

Nota. Af: Afabilidad; Ap: Apertura; Es: Estabilidad emocional; Ex: Extraversión; Re: Responsabilidad; +: ítem directo o polaridad positiva; -: ítem inverso o polaridad negativa.

puede no eliminarse (Brown & Maydeu-Olivares, 2018). Algunos meta-análisis muestran que las pruebas cuasi-ipsativas tienen mayor validez predictiva (Salgado et al., 2015; Salgado & Táuriz, 2014) y son más robustas al falseamiento (Martínez & Salgado, 2021).

Los cuatro formatos más frecuentes de las pruebas de elección forzosa son (ver Tabla 1): (a) elegir el ítem que mejor te describe de entre dos enunciados (PICK-PAIR), (b) elegir el ítem que mejor te describe de entre más de dos enunciados (PICK), (c) elegir el ítem que más te describe y el que menos (MOLE, de "MOst and LEast"), y (d) ordenar las alternativas según el grado en el que te describen (RANK). En cuanto a la puntuación tradicional, en los formatos PICK y PICK-PAIR se puede puntuar +1 en la dimensión si la polaridad del ítem elegido es positiva (ver Tabla 1) o -1, si fuera negativa (i.e., ítem inverso). En el formato RANK se pueden asignar valores entre 1 y *K*, siendo *K* el número de frases a ordenar, mientras que en el formato MOLE se pueden asignar puntuaciones -1, 0 o 1, dependiendo de la elección concreta y la polaridad de los ítem seleccionados (ver dos ejemplos en Tabla 1). Hontangas et al. (2015, 2016) encontraron, por simulación, que el formato MOLE proporcionaba resultados similares al RANK y, en ambos casos, superiores al PICK. Sin embargo, Cao y Drasgow (2019) encuentran que el formato PICK es más resistente al falseamiento que el formato MOLE, indicando que este último, además, implica una mayor carga cognitiva para responder.

PUNTUANDO BLOQUES DE ELECCIÓN FORZOSA DESDE LA TRI

En los últimos años, se ha sugerido que muchos de los problemas de las puntuaciones en pruebas de elección forzosa pueden deberse al propio procedimiento clásico de puntuación, pudiendo superarse mediante el modelado de las respuestas desde la Teoría de la Respuesta al Ítem (TRI; p.ej., Brown & Maydeu-Olivares, 2011; Hontangas et al., 2015, 2016; Morillo et al., 2016). La TRI permite modelar las probabilidades de respuesta a los bloques en función de los niveles de rasgo, lo que posibilita alcanzar, bajo ciertas condiciones, una interpretación normativa de las puntuaciones (i.e., posibilita comparaciones entre individuos). El uso de modelos de TRI viene acompañado de diversas ventajas (Olea et al., 2010): (a) permite evaluar la precisión para cada nivel de rasgo, en vez de asumir que todas las personas son evaluadas con la misma fiabilidad; (b) permite obtener puntaciones en la misma escala, aun cuando se apliquen distintos ítems; y (c) permite el desarrollo de aplicaciones avanzadas, como los Test Adaptativos Informatizados (TAIs). La principal característica de los TAIs es que los ítems administrados se ajustan al nivel de rasgo que va manifestando el evaluado según sus respuestas a los ítems previos. El uso de un TAI permite obtener medidas más eficientes (igual precisión en menor tiempo), así como medidas con un nivel de precisión más homogéneo a través del nivel de rasgo.

Se han propuesto diversos tipos de modelos de TRI para describir el proceso de comparación de ítems dentro de un bloque, de entre los que destacan el MUPP (Multi-unidimensional Pairwise Preference) y el TIRT (Thurstonian Item Response Theory).

El modelo MUPP fue desarrollado por Stark et al. (2005) para bloques de dos ítems, cada uno midiendo una dimensión distinta. En primer lugar, se define un modelo para la probabilidad de que una persona esté de acuerdo con el contenido de un ítem. Puede asumirse que esta probabilidad sigue un modelo de dominancia (Morillo et al., 2016) o un modelo de punto ideal (Stark et al., 2005). En un modelo de dominancia la probabilidad de acuerdo con un ítem (p.ej., "Creo que los demás tienen buenas intenciones") aumenta en función del nivel de rasgo (p.ej., la Afabilidad). En un modelo de punto ideal la función de probabilidad de respuesta es unimodal; esto es, la probabilidad de acuerdo aumenta en función del nivel de rasgo hasta alcanzar un máximo y luego se reduce. Por ejemplo, la probabilidad de acuerdo con el ítem "A veces puedo persuadir a mis amigos de que hagan las cosas a mi manera" puede ser máxima para personas que tienen una cierta capacidad de persuasión, pero menor para personas que nunca persuaden a sus amigos o para personas que siempre les persuaden. En segundo lugar, a partir de estas probabilidades de acuerdo con los ítems, puede obtenerse la probabilidad de preferir un ítem sobre otro dentro de un bloque (ver, por ejemplo, Morillo et al., 2016).

El modelo TIRT (Brown & Maydeu-Olivares, 2011) se basa en la ley del juicio comparativo de Thurstone y asume un modelo de dominancia. En este modelo se desglosa la respuesta a cada bloque en un conjunto de comparaciones binarias. Por ejemplo, supongamos que alguien establece que, en un bloque de tres ítems, la frase que más le representa es la B y la que menos la A. Ese ordenamiento (B > C > A) se podría representar con tres variables, una por comparación binaria: $X_{AB} = 0$ (i.e., prefiere el ítem B al A), $X_{BC} = 1$ (i.e., prefiere el ítem B al C) y $X_{AC} = 0$ (i.e., prefiere el ítem C al A). Una vez creadas estas variables pueden estimarse los modelos de TRI mediante análisis factorial (Brown & Maydeu-Olivares, 2012). En el caso de bloques de dos enunciados, Morillo et al. (2016) muestran que, cuando se asume el modelo de dominancia, el MUPP es equivalente al TIRT.

¿Qué modelo es mejor? Los modelos MUPP de punto ideal son más flexibles, pero quizás innecesariamente complejos. La cuestión decisiva se encontraría en aceptar o no la necesidad de usar ítems con funciones de probabilidad unimodales. Ítems como "A veces puedo persuadir a mis amigos de que hagan las cosas a mi manera" suelen desecharse en el análisis psicométrico previo y con frecuencia son ambiguos e incluso frustrantes para los respondientes (Brown & Maydeu-Olivares, 2010). No obstante, algunos de los mayores éxitos en la aplicación de la TRI a ítems de elección forzosa se han conseguido con modelos de punto ideal.

En cualquier caso, la ventaja de usar modelos de TRI radica en obtener puntuaciones con menor, o incluso nula ipsatividad, aunque el grado en que esto se consigue dependerá del diseño del test, como se describe a continuación.

FACTORES GENERALES QUE AFECTAN A LA EFICACIA DE LOS TEST DE ELECCIÓN FORZOSA

Como se ha mencionado, los modelos de TRI no dan necesariamente lugar a puntuaciones con propiedades normativas. Frick et al. (2021) señalan algunos factores que afectan a la eficacia en la construcción de ítems de elección forzosa en el caso de modelos de dominancia, aunque sus conclusiones no pueden considerarse definitivas. En primer lugar, su recomendación más importante es la de utilizar tanto bloques homopolares positivos, formados por ítems que

miden las dimensiones en la misma dirección (p.ej., A. *Creo que es emocionante hablar con muchas personas diferentes {Ex+}*; B. *Me siento cómodo conmigo mismo {Es+}*), como bloques heteropolares, formados por ítems que miden las dimensiones en direcciones contrarias (p.ej., A. *Me gusta hablar con extraños {Ex+}*; B. *Me preocupo por las cosas {Es-}*). Por ejemplo, si se utilizan únicamente bloques del primer tipo puede ser más difícil saber si alguien escoge el ítem A por tener alta extroversión o por tener baja estabilidad, mientras que incluir bloques del segundo tipo ayudará a distinguir esos dos perfiles. A pesar de esto, la necesidad de usar bloques heteropolares es discutible, ya que puede ser más difícil igualar los ítems en deseabilidad social (Bürkner et al., 2019; Lee & Joo, 2021), facilitando el falseamiento que se pretende prevenir en este formato. Por otro lado, Morillo et al. (2016) y Kreitchmann et al. (2021) han mostrado la posibilidad de estimar con precisión sin bloques heteropolares siempre que se haga un ensamblaje óptimo y haya un rango suficiente en los pesos de los ítems (algo que también apuntan Frick et al., 2021).

Otro factor importante es el tamaño de los bloques. Incrementar su tamaño (p.ej., usando tripletas) puede reducir la ipsatividad, pero incrementa la carga cognitiva al requerir más comparaciones por bloque (Sass et al., 2020). De hecho, Frick et al. (2021) encuentran fiabilidades similares cuando se comparan pares y tripletas si se mantiene constante el número de comparaciones binarias. Otro problema de los bloques de más de dos ítems es que, al aplicar el TIRT en ausencia de bloques heteropolares, se tiende a sobrestimar la fiabilidad.

Otros factores relevantes para la presencia de ipsatividad son las correlaciones entre dimensiones y el número de estas. Cuanto menor sea el número de dimensiones medidas o mayor la correlación positiva, mayor será la ipsatividad. Por ejemplo, a partir de los resultados en sus estudios de simulación, Bürkner et al. (2019) sugieren que con cinco o menos dimensiones y bloques homopolares no se pueden alcanzar mediciones precisas, mientras que con 30 sí obtienen una buena recuperación (no consideraron casos intermedios, entre 6 y 29 factores). Fisher et al., (2019) también son pesimistas en cuanto al uso del TIRT, encontrando peor validez referida a criterio en contextos de selección. En realidad, son varios los estudios empíricos que encuentran que las fiabilidades de las pruebas pueden ser bajas (p.ej., Kreitchmann et al., 2019) o que las correlaciones entre rasgos pueden verse distorsionadas (p.ej., Morillo et al., 2016). Probablemente, estas inconsistencias entre estudios se deben a la dificultad para construir buenos bloques homopolares de elección forzosa.

LA CONSTRUCCIÓN DE BLOQUES DE ELECCIÓN FORZOSA

La clave para el éxito en la construcción de una prueba de elección forzosa es el emparejamiento de los ítems en deseabilidad social, atendiendo a la información que proporciona el bloque en su conjunto. Respecto al emparejamiento en deseabilidad, suele recurrirse a valoraciones de expertos (o de muestras similares a la que es objeto de evaluación) para puntuar la deseabilidad social de los ítems. En este punto, Pavlov et al., (2021) destacan la importancia de emparejar los ítems no solo por deseabilidad social, sino teniendo en cuenta el consenso de los jueces en la valoración.

Respecto a la formación de bloques informativos, la utilización de un modelo de TRI permite anticipar cuánta información proporcionará el bloque cuando se aplique (esto es, el grado en que reducirá la varianza error de los niveles de rasgo estimados) y ensamblar ítems en bloques para maximizar la información. No obstante, un problema suele ser el tamaño del universo potencial de bloques. Por ejemplo, ensamblar 60 ítems en 30 bloques de 2 deriva, aproximadamente, en 2.92×10^{40} cuestionarios posibles (Kreitchmann et al., 2021). Para resolver este problema, Kreitchmann et al. (2021) adaptan el algoritmo genético NHBSA (node histogram-based sampling algorithm; Tsutsui, 2006) al problema de ensamblar ítems en bloques y proporcionan una implementación amigable en Shiny que permite diseñar una prueba de elección forzosa (https://psychometricmodelling.shinyapps.io/FCoptimization/). Kreitchmann et al. (2021) encontraron que el algoritmo propuesto era más eficiente que los métodos ya existentes (p.ej., al azar con restricciones de contenido o fuerza bruta). En resumen, la calidad de una prueba de elección forzosa dependería, como en test tradicionales, de la calidad psicométrica de sus componentes: los bloques que lo forman.

TEST ADAPTATIVOS INFORMATIZADOS EN PERSONALIDAD CON FORMATO LIKERT

En el ámbito de la personalidad pueden encontrarse algunos ejemplos de TAIs para medir los Big Five con escalas Likert. Destaca el trabajo pionero de Reise y Henson (2000) para el NEO-PI-R en el que encontraron que un TAI-unidimensional de tan solo cuatro ítems por faceta (esto es, reduciendo la longitud a la mitad) proporcionaba una buena recuperación de los niveles de rasgo. También se han desarrollado TAIs basados en modelos multidimensionales asumiendo factores correlacionados (p.ej., Makransky et al., 2013; Nieto et al., 2018) y basados en el modelo bifactor (Nieto et al., 2018), aplicados dentro de cada dominio de personalidad (p.ej., Extraversión). Los TAIs multidimensionales muestran cierta ventaja al tener en cuenta las correlaciones entre las distintas facetas (p.ej., en el estudio de Makransky et al., 2013, se obtuvo una correlación promedio elevada de 0,7 para las facetas del dominio de estabilidad emocional). En los estudios de Nieto et al. (2017; 2018) se investigaron las correlaciones entre las puntuaciones obtenidas en los TAI con las obtenidas en el banco completo. Para los dominios, con 12 ítems por dominio, se alcanzaban correlaciones promedio de 0,89 para el TAI unidimensional (y para las escalas cortas), y de 0,94 para los TAI multidimensionales (Nieto et al., 2018). Estos últimos, además, proporcionaban un mejor balance en la proporción de ítems aplicados en cada faceta (i.e., mayor validez de contenido). Para las facetas, los TAIs multidimensionales alcanzaban una correlación promedio más baja que los unidimensionales (0,87 vs. 0,95), pero con la mitad de los ítems.

CONSTRUYENDO TAIS DE ELECCIÓN FORZOSA ADAPTATIVOS

Las ventajas de un TAI pueden hacerse especialmente importantes en ítems de pocas categorías de respuesta, como en el formato PICK-PAIR, ya que en esos casos el rango de niveles de rasgo para los que el ítem es preciso es estrecho. Existen múltiples TAIs de elección forzosa (TAI-EF), siendo el más famoso el TAPAS (p.ej., Stark et al., 2014), que mide 22 dimensiones de personalidad y que se en-

tiende como un "test a la carta", pudiéndose elegir, por ejemplo, las dimensiones a evaluar, el tipo de test (adaptativo o fijo) y el formato (p.ej., binario, politómico, de elección forzosa unidimensional o de elección forzosa multidimensional) en función del contexto de aplicación (p.ej., mayor o menor previsión de deseabilidad social). Las versiones adaptativas permiten reducir la longitud a la mitad (Drasgow et al., 2012).

La efectividad de un TAI-EF multidimensional depende de: (a) el banco de bloques ensamblado y (b) la regla de selección. Respecto al primer punto, lo mencionado en apartados anteriores para la construcción de test fijos es aplicable. Los bloques pueden emparejarse, por ejemplo, atendiendo a un algoritmo genético, produciendo bancos óptimos de los que seleccionar bloques adaptativamente. En cuanto a la regla de selección, existen distintas variantes. En modelos unidimensionales, la varianza error es inversamente proporcional a la información del test, que es la suma de las funciones de información de los ítems. Igualmente, la matriz de varianzas-covarianzas error en modelos multidimensionales es la inversa de la matriz de información. A pesar de la aparente similitud, esta diferencia implica que distintas reglas (p.ej., regla-T: maximizar la información de cada dimensión al añadir el ítem; regla-A: minimizar la varianza error de cada dimensión al añadir el ítem) dan distintos resultados.

Por ejemplo, Kreitchmann et al. (enviado) parten de un banco de 240 ítems (48 ítems por dimensión) que ensamblan en un banco de 120 bloques. El número de bloques posibles, excluyendo los unidimensionales, era de 23.040. En este caso, se compararon los resultados para TAIs de distinta longitud (i.e., 30 y 60 bloques) y regla de selección (p.ej., regla-T y regla-A), partiendo de un banco construido según el algoritmo genético o de un banco conformado por bloques al azar. Para la mejor regla de selección se encontró que, en promedio, el uso de un banco óptimo frente a un banco aleatorio podía incrementar el coeficiente de fiabilidad 0,05 puntos (de 0,80 a 0,85) y, más importante, reducir la ipsatividad de las puntuaciones (el sesgo negativo de las correlaciones entre dimensiones y en la relación con el criterio se reducía en 0,04 puntos). En cuanto a la regla de selección, Kreitchmann et al. (enviado) encontraron que, en consonancia con investigaciones previas (p.ej., Mulder & van der Linden, 2009), la regla-A, al minimizar directamente las varianzas de error, proporcionaba mejores resultados. Este resultado es importante, puesto que algunos investigadores utilizan la regla-T por eficiencia computacional (Chen et al., 2020).

CONSTRUYENDO TAIS DE ELECCIÓN FORZOSA ADAPTATIVOS ON-THE-FLY

Como se ha mencionado, el uso de un algoritmo genético permite la optimización de un test fijo o de un banco. El siguiente paso natural es el de construir bloques on-the-fly; esto es, ensamblar "al vuelo" los enunciados en bloques en el momento de aplicar el TAI. Esta idea constituye el corazón del TAPAS, que parte de un conjunto amplio de ítems (enunciados) calibrados de los que se derivan, mediante el modelo MUPP, un conjunto gigante de bloques potenciales de los que se selecciona en cada momento el más informativo. Este proceder no está exento de supuestos, ya que asume la veracidad del modelo MUPP y la ausencia de efectos del contexto (i.e., el funcionamiento de cada ítem no depende del ítem con el que se empareje). Aunque puede haber efectos del contexto, este supuesto de invarianza se puede sostener razonablemente en la práctica (Lin & Brown, 2017; Morillo et al., 2019). Lin y Brown (2017) sugieren que los efectos de contexto pueden reducirse si se emparejan los ítems en deseabilidad social (de lo contrario, el ítem claramente más deseable será más elegido por ser percibido como la "respuesta correcta") e indican que debe evitarse incluir dentro del mismo bloque ítems similares en contenido (p.ej., *Soy una persona animada en la conversación; Evito hablar de mis éxitos*), ya que esto puede modificar el significado de los ítems (en el ejemplo, *Evito hablar de mis éxitos* dejaría de ser un marcador de modestia para ser un marcador de extroversión). En todo caso, los resultados de validez predictiva del TAPAS, cuya prueba adaptativa se basa en esa invarianza, son positivos (p.ej., Trent et al., 2020).

Kreitchmann et al. (enviado) encuentran que un TAI-EF on-the-fly, bajo ese supuesto de invarianza y un procedimiento de selección óptimo, muestra una pequeña mejora respecto a un TAI-EF basado en un banco óptimo (p.ej., 0,01 en el coeficiente de fiabilidad), pero grandes mejoras en el control de la exposición, ya que al incrementarse el número de bloques posibles es más difícil que dos evaluados reciban exactamente los mismos bloques.

DISCUSIÓN

El avance de la tecnología y el desarrollo de modelos psicométricos en las dos últimas décadas está permitiendo dar respuesta a un problema clásico: la medición de la personalidad en contextos de selección en los que la deseabilidad social puede ser alta. En 2005, Salgado incluía las pruebas de elección forzosa como una solución no recomendable, en parte por la dificultad de análisis que sufre este tipo de pruebas. La evidencia sobre su mayor robustez al falseamiento y mayor validez predictiva parece inclinar la balanza hacia una visión más positiva, siempre que se resuelvan los problemas de ipsatividad de las puntuaciones. Desde la TRI, se han propuesto distintos modelos (Brown & Maydeu-Olivares, 2011; Morillo et al., 2016; Stark et al., 2005) que ayudan a resolver el problema de la ipsatividad. No obstante, las demandas de evaluación, el ensamblaje de los bloques y el tamaño de los bloques pueden tener gran importancia. En general, el comportamiento de las pruebas será mejor cuantas más dimensiones se midan y menos correlacionadas estén. En el ensamblaje de bloques es relevante no sólo el emparejamiento por deseabilidad social, sino también su contribución a reducir la varianza error. A este respecto, la necesidad de usar bloques heteropolares para eliminar la ipsatividad de las puntuaciones (Frick et al., 2021) o no (Morillo, 2018; Kreitchmann et al., 2021) es un tema de debate. Para Bürkner et al. (2019), los bloques heteropolares podrían ser contraproducente en contextos aplicados. Nuestra recomendación es el uso de algoritmos de optimización para conformar los bloques homopolares óptimamente. Esto puede resultar costoso, ya que requiere recoger información sobre la deseabilidad de los ítems, así como su calibración previa. No obstante, no debemos olvidar que: (a) el análisis exploratorio de la estructura puede resultar más complejo *a posteriori,* en bloques bidimensionales; y (b) el ensamblaje no óptimo dará lugar a problemas de ipsatividad de las pun-

tuaciones. Por último, el tamaño de los bloques introduce una complejidad añadida en la creación de bloques óptimos puesto que, a medida que se incrementa el tamaño de los bloques, el número de bloques posibles entre los que elegir incrementa exponencialmente, haciendo menos factible explorar el universo de posibilidades.

En definitiva, aunque las pruebas de elección forzosa existen desde mucho tiempo atrás su uso se ha visto reducido por las limitaciones que se les atribuían. Sin embargo, cada vez se entiende mejor que este tipo de pruebas no constituye una categoría homogénea, siendo importante comprender cómo el diseño de la prueba y el modo de puntuación influye en su robustez al falseamiento, en la resolución del problema de la ipsatividad y, en definitiva, en su validez predictiva. Los meta-análisis más recientes muestran que, en contextos aplicados, el uso de pruebas de elección forzosa cuasi-ipsativas constituye una estrategia prometedora para obtener una mayor validez predictiva en el ámbito de la personalidad, con mayor resistencia al falseamiento que otros formatos (Martinez y Salgado, 2021). Finalmente, el avance de las nuevas tecnologías y el desarrollo de nuevos modelos psicométricos se presentan como dos potentes aliados que posibilitan la construcción de pruebas adaptadas al vuelo, optimizando el diseño de las pruebas y la puntuación de estas.

CONFLICTO DE INTERESES

No existe conflicto de intereses

REFERENCIAS

Abad, F. J., Sorrel, M. A., Garcia, L. F., & Aluja, A. (2018). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment, 25*(8), 959-977. https://doi.org/10.1177/1073191116667547

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A Meta-Analysis. *Personnel Psychology, 44*(1), 1-26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460-502. https://doi.org/10.1177%2F0013164410375112

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology, 3*(4), 489-493. https://doi.org/10.1111/j.1754-9434.2010.01277.x

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior research methods, 44*(4), 1135-1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2018). Modelling forced-choice response formats. En P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 523-569). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489772.ch18

Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 79*(5), 827-854. https://doi.org/10.1177/0013164419832063

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347-1368. https://doi.org/10.1037/apl0000414

Chen, C., Wang, W., Chiu, M. M., & Ro, S. (2020). Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items. *Journal of Educational Measurement, 57*(2), 343-369. https://doi.org/10.1111/jedm.12252

Cuadrado, D., Salgado, J. F., & Moscoso, S. (2021). Personality, intelligence, and counterproductive academic behaviors: A meta-analysis. *Journal of Personality and Social Psychology, 120*(2), 504-537. https://doi.org/10.1037/pspp0000285

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army personnel selection and classification decisions*. Drasgow Consulting Group Urbana IL.

Fisher, P., Robie, C., Christiansen, N., Speer, A., & Schneider, L. (2019). Criterion-related validity of forced-choice personality measures: A cautionary note regarding Thurstonian IRT versus classical test theory scoring. *Personnel Assessment and Decisions, 5*(1). https://doi.org/10.25035/pad.2019.01.003

Frick, S., Brown, A. & Eunike Wetzel (2021) Investigating the normativity of trait estimates from multidimensional forced-choice data, *Multivariate Behavioral Research*, https://doi.org/10.1080/00273171.2021.1938960

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9-24. https://doi.org/10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167-184. https://doi.org/10.1037/h0029780

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement, 39*(8), 598-612. https://doi.org/10.1177/0146621615585851

Hontangas, P. M., Leenen, I., & de la Torre, J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema, 28*, 1, 76-82. https://doi.org/10.7334/psicothema2015.204

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*(6), 875-925. https://doi.org/10.1037/a0033901

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items. *Frontiers in Psychology, 10*, 2309. https://doi.org/10.3389/fpsyg.2019.02309

Kreitchmann, R. S., Abad, F. J., & Sorrel, M. A. (2021). A genetic algorithm for optimal assembly of pairwise forced-choice questionnaires. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01677-4

Kreitchmann, R. S., Sorrel, M. A., & Abad, F. J. (enviado). On bank

assembly and block selection in multidimensional forced-choice adaptive assessments.

Lee, P., & Joo, S.-H. (2021). A new investigation of fake resistance of a multidimensional forced-choice measure: An application of differential item/test functioning. *Personnel Assessment and Decisions, 7*(1). https://doi.org/10.25035/pad.2021.01.004

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement, 77*(3), 389-414. https://doi.org/10.1177%2F0013164416646162

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the Neo Pi-R. *Assessment, 20*(1), 3-13. https://doi.org/10.1177/1073191112437756

Martínez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology, 12*, 732241. https://doi.org/10.3389/fpsyg.2021.732241

Morillo, D. (2018). *Item response theory models for forced-choice questionnaires*. Doctoral dissertation, Universidad Autónoma de, Madrid.

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Journal of Work and Organizational Psychology, 35*(2), 75-83. https://doi.org/10.5093/jwop2019a11

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement, 40*(7), 500-516. https://doi.org/10.1177/0146621616662226

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*(2), 273-296. https://doi.org/10.1007/s11336-008-9097-5

Nieto, M. D., Abad, F. J., & Hernández-Camacho, A. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29*.3, 390-395. https://doi.org/10.7334/psicothema2016.391

Nieto, M. D., Abad, F. J., & Olea, J. (2018). Assessing the Big Five with bifactor computerized adaptive testing. Psychological *Assessment, 30*(12), 1678-1690. https://doi.org/10.1037/pas0000631

Olea, J., Abad, F. J., & Barrada, J. R. (2010). Tests informatizados y otros nuevos tipos de tests. *Papeles del psicólogo, 31*(1), 97-107.

Otero, I., Cuadrado, D., & Martínez, A. (2020). Convergent and predictive validity of the Big Five factors assessed with single stimulus and cuasi-ipsative questionnaires. *Journal of Work and Organizational Psychology, 36*(3), 215-222. https://doi.org/10.5093/jwop2020a17

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences, 183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. https://doi.org/10.1037/a0014996

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*(4), 347-364. https://doi.org/10.1177%2F107319110000700404

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387. https://doi.org/10.1037/a0026838

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*, 115-128.

Salgado, J. F. (2016). A theoretical model of psychometric effects of faking on assessment procedures: Empirical findings and implications for personality at work: A Theoretical Model of faking psychometric effects. *International Journal of Selection and Assessment, 24*(3), 209-228. https://doi.org/10.1111/ijsa.12142

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and cuasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 88*(4), 797-834. https://doi.org/10.1111/joop.12098

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3-30. https://doi.org/10.1080/1359432X.2012.716198

Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment, 27*(3), 572-584. https://doi.org/10.1177/1073191118762049

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184-203. https://doi.org/10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification secisions. *Military Psychology, 26*(3), 153-164. https://doi.org/10.1037/mil0000044

Trent, J. D., Barron, L. G., Rose, M. R., & Carretta, T. R. (2020). Tailored Adaptive Personality Assessment System (TAPAS) as an indicator for counterproductive work behavior: Comparing validity in applicant, honest, and directed faking conditions. *Military Psychology, 32*(1), 51-59. https://doi.org/10.1080/08995605.2019.1652481

Tsutsui, S. (2006). Node histogram vs. edge histogram: A comparison of probabilistic model-building genetic algorithms in permutation domains. 2006 *IEEE International Conference on Evolutionary Computation*, 1939-1946. https://doi.org/10.1109/CEC.2006.1688a44

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197-210. https://doi.org/10.1177/00131649921969802