PRUEBAS DE ELECCIÓN FORZOSA: VISIÓN ACTUAL Y RECOMENDACIONES

FORCED-CHOICE TESTS: CURRENT PERSPECTIVE AND RECOMMENDATIONS

FRANCISCO J. ABAD[1], RODRIGO S. KREITCHMANN[2], DIEGO F. GRAÑA[1], PABLO NÁJERA[3] Y MIGUEL A. SORREL[1]

Cómo referenciar este artículo/How to reference this article:

Abad, F. J., Kreitchmann, R. S., Graña, D. F., Nájera, P. y Sorrel, M. A. (2024). Pruebas de Elección Forzosa: Visión Actual y Recomendaciones [Forced-Choice Tests: Current Perspective and Recommendations]. *Acción Psicológica, 22*(1), 57–72. https://doi.org/10.5944/ap.22.1.43413

Resumen

Este artículo tiene como objetivo ofrecer una visión actual de las pruebas de elección forzosa y proporcionar recomendaciones para su diseño y construcción. Aunque estas pruebas ayudan a superar limitaciones como los sesgos de deseabilidad social y de respuesta extrema, comunes en los formatos de respuesta graduada, presentan desafíos técnicos relacionados con la ipsatividad de las puntuaciones. Este artículo presenta modelos psicométricos basados en la Teoría de Respuesta al Ítem (TRI), como el modelo Thurstoniano de TRI para preferencias (TIRT) y el modelo de preferencia por pares multi-unidimensional (MUPP), que mejoran la estimación de los rasgos y permiten un ensamblaje óptimo de ítems en bloques. Se identifican factores de diseño del cuestionario, como la polaridad de los ítems ensamblados, que pueden afectar la calidad de las puntuaciones obtenidas. Además, se exploran los beneficios de la TRI en el desarrollo de tests adaptativos informatizados *on-the-fly*, donde los ítems se emparejan durante la prueba en

Correspondence address [Dirección para correspondencia]: Rodrigo S. Kreitchmann. Facultad de Psicología. Universidad Nacional de Educación a Distancia, Madrid, España.
Email: rschames@psi.uned.es

ORCID: Francisco J. Abad (https://orcid.org/0000-0001-6728-2709), Rodrigo S. Kreitchmann (https://orcid.org/0000-0001-5199-9828), Diego F. Graña (https://orcid.org/0009-0005-2198-5341), Pablo Nájera (https://orcid.org/0000-0001-7435-2744) y Miguel A. Sorrel (https://orcid.org/0000-0002-5234-5217).

[1] Universidad Autónoma de Madrid, España.
[2] Universidad Nacional de Educación a Distancia, España.
[3] Universidad Pontificia Comillas, España.

Agradecimientos: Este trabajo ha sido financiado por MICIU/AEI/10.13039/501100011033 y FEDER, UE (proyecto PID2022-137258NB-I00), por la Cátedra de Modelos y Aplicaciones Psicométricos (Instituto de Ingeniería del Conocimiento y Universidad Autónoma de Madrid) y por la ayuda PREP2022-001047, financiada por MICIU/AEI/10.13039/501100011033 y el FSE+.

Recibido: 18 de noviembre de 2024.
Aceptado: 23 de febrero de 2025.

función de las respuestas previas del evaluado, optimizando la precisión de las puntuaciones. Finalmente, se ofrece una guía paso a paso para la construcción de pruebas de elección forzosa, ilustrada con un ejemplo empírico y código en R de acceso abierto.

Palabras clave: Personalidad; Elección forzosa; Test adaptativos informatizados; *On-the-Fly*.

Abstract

This article aims to provide a current overview of forced-choice tests and offer recommendations for their design and construction. Although these tests help overcome limitations such as social desirability and extreme response bias, common in graded response formats, they present technical challenges related to the ipsativity of the scores. This paper discusses psychometric models based on Item Response Theory (IRT), such as the Thurstonian IRT (TIRT) and the multi-unidimensional pairwise preference (MUPP) models, which improve trait estimation and enable optimal item assembly into blocks. It identifies questionnaire design factors, such as the polarity of assembled items, that can affect the quality of the obtained scores. Additionally, the benefits of IRT in developing computerized adaptive tests on-the-fly are explored, where items are paired during the test based on the examinee's previous responses, optimizing score precision. Finally, a step-by-step guide for constructing forced-choice tests is provided, illustrated with an empirical example and open-access R code.

Keywords: Personality; Forced-choice; Computerized Adaptive Testing; On-the-fly.

Pruebas de elección forzosa: visión actual y recomendaciones

Aunque el origen y el debate sobre las pruebas de elección forzosa se remonta a mucho tiempo atrás (Zavala, 1965), su uso ha ido creciendo en popularidad en el campo de la evaluación psicométrica de la personalidad, especialmente en contextos laborales y educativos (Heggestad et al., 2006). Esto se debe a los problemas tradicionales de las pruebas de autoinforme que emplean un formato de respuesta de tipo Likert, tales como su menor robustez al falseamiento y la presencia de sesgos de respuesta como la deseabilidad social, aquiescencia o respuesta extrema (Cao y Drasgow, 2019; Kreitchmann et al., 2019). En términos generales, las pruebas de elección forzosa son herramientas efectivas para minimizar sesgos de respuesta y extraer inferencias precisas sobre atributos no cognitivos, como actitudes, valores, intereses vocacionales, motivaciones, competencias o estilos de aprendizaje, a partir de las preferencias relativas expresadas por los evaluados (véase Hontangas et al., 2015, para ejemplos específicos). No obstante, la construcción de una prueba de elección forzosa implica algunos desafíos técnicos que requieren el uso de modelos psicométricos avanzados en todo el proceso de desarrollo del test, desde el diseño y ensamblaje de los ítems (estímulos) en bloques, hasta la estimación del nivel de rasgo. En el presente trabajo, se recogen algu-nas de las herramientas y recomendaciones para el diseño óptimo de una prueba de elección forzosa, de forma que éstas puedan ofrecer puntuaciones fiables y válidas.

Este trabajo complementa y amplía investigaciones previas sobre el modelado de pruebas de elección forzosa (e.g., Abad et al., 2022). Mientras que dichos estudios se han centrado en la conceptualización general y los desafíos técnicos asociados a este tipo de pruebas, el presente manuscrito adopta un enfoque más aplicado. En particular, se presenta un tutorial detallado que guía al lector en la construcción y calibración de pruebas de elección forzosa, con un énfasis especial en los test fijos (frente a los test adaptativos).

Las pruebas de elección forzosa y el problema de la ipsatividad

El formato de elección forzosa se caracteriza por la presentación de bloques de dos o más enunciados, entre los que el evaluado debe indicar cuál le representa mejor, o establecer un ordenamiento, total o parcial, de estos (para una revisión completa, véase Brown y Maydeu-Olivares, 2018a; Hontangas et al., 2015, 2016). En el caso

más simple, el formato binario (PICK-PAIR), se pide a la persona que seleccione el ítem que mejor le describa de entre dos enunciados (e.g., "Me gusta innovar en lo que hago" y "Me considero una persona feliz"). Otros formatos comunes consisten en pedir a la persona que escoja el ítem que mejor le represente entre más de dos enunciados (PICK), elegir el ítem que más y el que menos le describe (MOLE, de "MOst and LEast"), u ordenar las opciones según el grado en que su descripción le representa (RANK). Respecto a la puntuación tradicional bajo la teoría clásica de los tests, en los formatos PICK y PICK-PAIR se otorga +1 a la dimensión si el enunciado escogido posee polaridad positiva (mide directamente la dimensión) o -1 si tiene polaridad negativa (mide la dimensión inversa). En el formato RANK se conceden valores que varían entre 1 y K, siendo K el número de enunciados. Por otro lado, en el sistema MOLE, se asignan las puntuaciones -1, 0 o 1, dependiendo de la selección particular y la polaridad de los ítems escogidos.

Aunque los estudios de metaanálisis han mostrado la mayor robustez de este formato al falseamiento (Cao y Drasgow, 2019; Martínez y Salgado, 2021), el problema principal de estos cuestionarios es que las puntuaciones obtenidas tendrán propiedades *ipsativas* en mayor o menor grado (Hicks, 1970). Esto quiere decir que cada puntuación de un evaluado depende en parte de sus otras puntuaciones, pues manifiesta una predominancia relativa y no absoluta de los rasgos. Por ejemplo, una persona altamente organizada y algo sociable y otra menos organizada y poco sociable podrían coincidir en sus respuestas y puntuaciones ya que ambas se perciben como más organizadas que sociables. De forma similar, una tercera persona poco sociable y nada organizada podría obtener una puntuación en sociabilidad superior a la primera persona ya que las puntuaciones reflejan preferencias relativas por los rasgos. En el caso extremo de puntuaciones totalmente ipsativas, la suma de las puntuaciones de cada individuo será un valor constante e igual para todos los evaluados, lo que imposibilita realizar comparaciones entre sujetos, como deducir que la primera persona es más organizada que la segunda y que la tercera.

La Figura 1 muestra el problema de la ipsatividad en un modelo de puntuación tradicional para el formato PICK-PAIR, con bloques de dos ítems. Se presentan las respuestas de dos evaluados en bloques que miden tres rasgos de personalidad, donde todos los ítems tienen polaridad positiva y cada bloque evalúa rasgos distintos. Suponiendo conocer los niveles de rasgo verdaderos, el primer evaluado tiene puntuaciones típicas de 1,0 en estabilidad emocional (EE), 0,5 en extroversión (EX) y 0,0 en apertura a experiencias (AP). Por lo tanto, prefiere los ítems de EE en los bloques 1 y 2, y en el bloque 3, que mide EX y AP, elige el ítem de EX. Así, su puntuación tradicional es 2 en EE, 1 en EX y 0 en AP. El segundo evaluado tiene

Figura 1

Ejemplo de elección forzosa con bloques de dos ítems para dos evaluados ficticios

Evaluado 1

Nivel de Rasgo Verdadero:		
Estabilidad Emocional (EE)	Extroversión (EX)	Apertura (AP)
1,0	0,5	0,0
Respuestas:		
Rara vez me irrito. (EE) ✓		Hago amigos con facilidad. (EX) ○
Me gusta la innovación. (AP) ○		Me siento cómodo conmigo mismo. (EE) ✓
Sé cómo cautivar a la gente. (EX) ✓		Disfruto escuchando ideas nuevas. (AP) ○
Puntuación Ipsativa:		
Estabilidad Emocional (EE)	Extroversión (EX)	Apertura (AP)
2	1	0

Evaluado 2

Nivel de Rasgo Verdadero:		
Estabilidad Emocional (EE)	Extroversión (EX)	Apertura (AP)
-1,5	-1,0	-0,5
Respuestas:		
Rara vez me irrito. (EE) ○		Hago amigos con facilidad. (EX) ✓
Me gusta la innovación. (AP) ✓		Me siento cómodo conmigo mismo. (EE) ○
Sé cómo cautivar a la gente. (EX) ○		Disfruto escuchando ideas nuevas. (AP) ✓
Puntuación Ipsativa:		
Estabilidad Emocional (EE)	Extroversión (EX)	Apertura (AP)
0	1	2

puntuaciones típicas más bajas: -1,5 en EE, -1,0 en EX y -0,5 en AP. Su patrón refleja preferencia por los ítems de AP en los bloques 2 y 3, y por EX en el bloque 1, obteniendo una puntuación tradicional de 2 en AP, 1 en EX y 0 en EE. Este ejemplo evidencia que las puntuaciones ipsativas, aunque permiten medir la predominancia de rasgos en cada evaluado, no permiten comparaciones válidas entre ellos. Por ejemplo, aunque el primer evaluado tiene una puntuación típica verdadera más alta en AP, su puntuación ipsativa es inferior porque sus otros rasgos son más predominantes. En una muestra completa, las puntuaciones ipsativas en una dimensión mostrarán una covarianza negativa con las demás. Esto ocurre porque elegir más ítems de un rasgo implica seleccionar menos de los otros, generando interdependencia entre puntuaciones.

La ipsatividad plantea una serie de problemas: (a) distorsión de la dimensionalidad y estructura factorial del test (e.g., correlaciones entre dimensiones negativamente sesgadas); (b) sesgo en la validez predictiva (e.g., las correlaciones de las escalas con un criterio externo estarán sesgadas hacia cero); c) sesgos en los coeficientes de fiabilidad. Estos resultados se producen por las covarianzas negativas entre las puntuaciones de rasgo que surgen al hacer que los evaluados seleccionen una opción frente a otra de distinto rasgo. No obstante, estos problemas no son insalvables, ya que en los últimos años se han producido avances significativos tanto en el uso de modelos psicométricos que optimizan la puntuación de las pruebas como, de igual importancia, en el diseño de estas. Son numerosos los estudios que muestran que el formato de elección forzosa puede mantener propiedades psicométricas comparables a las del formato de respuesta graduada (e.g., Zhang et al., 2020) sugiriendo que, aunque estas medidas presentan limitaciones, pueden producir puntuaciones fiables y válidas. En las siguientes secciones se resume el conocimiento actual sobre ambos aspectos, haciendo finalmente recomendaciones sobre los pasos a seguir en la construcción de pruebas de elección forzosa.

Modelos psicométricos para pruebas de elección forzosa

De forma resumida, la ipsatividad ocurre cuando la elección de ítems de diferentes dimensiones da lugar a incrementos equivalentes en las puntuaciones de dichas dimensiones. Por ejemplo, si al seleccionar un ítem de extroversión incrementa la puntuación en dicho rasgo en la misma medida que elegir un ítem de responsabilidad incrementa la suya, y ocurre lo mismo para todos los ítems del cuestionario. El modelado psicométrico de respuestas de elección forzosa permite estimar parámetros de discriminación, capturando matices en la relación entre rasgos e ítems. Esto reduce la ipsatividad, ya que la relación entre la elección de un ítem y la puntuación en su dimensión no tiene por qué ser uniforme. Entre los modelos TRI más utilizados para este propósito destacan el TIRT (*Thurstone IRT*; Brown y Maydeu-Olivares, 2011) y el MUPP (*Multi-Unidimensional Pairwise-Preference*; Stark et al., 2005).

TIRT. El TIRT se basa en la ley de juicio comparativo de Thurstone, siendo un modelo para ítems que siguen un modelo de dominancia (i.e., la probabilidad de estar de acuerdo con un ítem sigue una relación monótona, creciente o decreciente, con el nivel de rasgo). Por ejemplo, la probabilidad de estar de acuerdo con "Soy una persona ordenada" tenderá a aumentar a medida que aumenta el nivel en el rasgo de responsabilidad. Desde este modelo, se descompone la respuesta al bloque de n ítems en términos de $n(n\text{-}1)/2$ respuestas en comparaciones binarias, modeladas a través de un modelo de TRI de ojiva normal. Para un bloque de tres ítems [*i, j, k*] cuya respuesta en formato RANK de un evaluado es $i = 1$ (mayor preferencia), $j = 2$, y $k = 3$ (menor preferencia), es posible inferir una preferencia por el ítem i en las comparaciones [*i, j*] e [*i, k*], y por el ítem j en la comparación [*j, k*]. Bajo el modelo TIRT se modelan conjuntamente estas pseudo-respuestas a las comparaciones entre pares. En la comparación [*i, j*], la probabilidad de elegir el elemento j se modelaría como:

$$P(j[i,j]|\boldsymbol{\theta}) = \Phi_N\left(\frac{(\boldsymbol{\lambda}_j' - \boldsymbol{\lambda}_i')\boldsymbol{\theta} - \gamma}{\sqrt{\psi_j^2 + \psi_i^2}}\right), \qquad [1]$$

donde Φ_N expresa la función de distribución normal acumulada, $\boldsymbol{\theta}$ es un vector con los niveles de rasgo, $\boldsymbol{\lambda_j' - \lambda_i'}$ es un vector que expresa la diferencia de pesos entre esos dos ítems del bloque, ψ_j^2 y ψ_i^2 reflejan las varianzas de especificidad, y γ es el parámetro de umbral para la compa-ración $[i, j]$, que se relacionaría con la dificultad para elegir el ítem j . La misma ecuación se generaliza para modelar las comparaciones $[i, k]$ y $[j, k]$.

En bloques de dos ítems unidimensionales, el modelo se simplifica a:

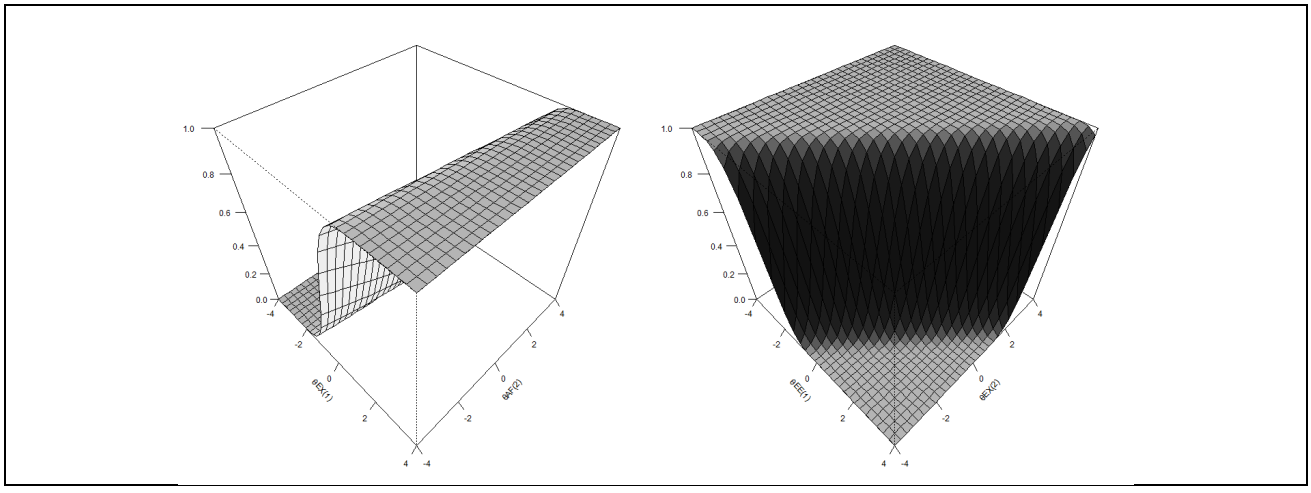
$$P(2[1,2]|\boldsymbol{\theta}) = \Phi_N\left(\frac{\lambda_{d2}\theta_{d2} - \lambda_{d1}\theta_{d1} - \gamma}{\sqrt{\psi_1^2 + \psi_2^2}}\right),$$

donde la probabilidad de elegir el ítem 2 del bloque se relaciona positiva (si el ítem 2 es directo) o negativamente (si el ítem 2 es inverso) con el nivel de rasgo medido por ese ítem (θ_{d2}), pero a la vez se relaciona positiva (si el ítem 1 es inverso) o negativamente (si el ítem 1 es directo) con el nivel de rasgo medido por el ítem 1 (θ_{d1}). En datos en los que los ítems no se repiten en diferentes bloques, con bloques binarios, los parámetros ψ_1^2 y ψ_2^2 no están identificados, por lo que la suma de ambos se fija a un valor. La Figura 2 muestra la representación de la superficie de res-puesta de dos bloques binarios, que siguen el modelo TIRT.

En el primer bloque, se representa la probabilidad de elegir "*Presiono a las personas*" (ítem 2) frente a "*Evito las multitudes*" (ítem 1). La probabilidad está relacionada con la dimensión del primer ítem, θ_{EX} (extraversión), pero no guarda relación con la dimensión del segundo ítem, θ_{AF} (afabilidad). Como el peso del primer ítem en θ_{EX} es nega-tivo, la relación a nivel de bloque resulta positiva: a mayor extraversión, menor es la probabilidad de elegir el ítem 1 y, por lo tanto, mayor probabilidad de elegir el ítem 2. En el segundo bloque, se representa la probabilidad de elegir "*Hago amigos fácilmente*" (ítem 2) frente a "*Mantengo la calma*" (ítem 1). En este caso, la probabilidad de elegir el segundo ítem depende de ambas dimensiones latentes, θ_{EX} y θ_{EE} (estabilidad emocional). El bloque tiene pesos simi-lares en las dos dimensiones y, dado que ambos ítems son directos (con pesos positivos), la probabilidad de elegir el ítem 2 (indicador de extroversión) aumenta con la dimen-sión θ_{EX} y disminuye con θ_{EE}. Además, se observa que es más fácil elegir el ítem 2 en el segundo bloque que en el primero (el umbral para elegir el ítem 2 del bloque, γ, es más

Figura 2

Representación de la superficie de respuesta en el TIRT, que representa la probabilidad de elegir el ítem 2 para un bloque con parámetros $\lambda_{EX(1)}$= -0.83, $\lambda_{AF(2)}$ = -0.09 y γ = .26 (izquierda) y para otro con parámetros $\lambda_{EE(1)}$ = 0.71, $\lambda_{EX(2)}$ = 0.61 y γ = -1.07 (derecha)



bajo). Esto sugiere que es más fácil afirmar la facili-dad para hacer amigos (frente a mantener la calma) que afirmar que se presiona a las personas (frente a evitar las multitudes).

MUPP. El modelo MUPP (Stark et al., 2005) establece que la probabilidad de elegir *j*, de entre *i*, *j* o *k*, sigue el axioma de Luce:

$$P(j[i,j,k]|\boldsymbol{\theta}) = \frac{P(j)\prod_{r\neq j} Q(r)}{\sum_s P(s)\prod_{r\neq s} Q(r)},$$

donde *P*(*j*) indica la probabilidad de valorar que el enunciado del bloque le representa al evaluado y *Q*(*r*) la probabilidad de valorar que el bloque *r* no le representa. La probabilidad *P*(*j*) puede modelarse por diferentes modelos bajo la TRI, como son los de dominancia o de punto-ideal. En los modelos de dominancia, como el modelo lo-gístico de dos parámetros (2PL), la probabilidad de res-puesta es monotónica creciente o decreciente con el nivel de rasgo. Por otro lado, en los modelos de punto ideal la probabilidad de acuerdo con un ítem es unimodal (i. e., máxima en un nivel de rasgo no extremo). Por ejemplo, la probabilidad de estar de acuerdo con "*Soy una persona medianamente ordenada*" puede ser máxima en personas con un nivel de bajas para niveles de rasgo más extremos (muy baja responsabilidad y nada de orden, o muy alta responsabilidad y mucho orden). Una reflexión sobre las ventajas relativas de cada modelo puede encontrarse en Brown y Maydeu-Olivares (2010). Para bloques de dos ítems, el modelo de dominancia da lugar al modelo MUPP-2PL (Morillo et al., 2016), donde el axioma de Luce puede simplificarse a:

$$P(2[1,2]|\boldsymbol{\theta}) = \psi_{logistico}\left((\mathbf{a'}_2 - \mathbf{a'}_1)\boldsymbol{\theta} + c\right), \qquad [2]$$

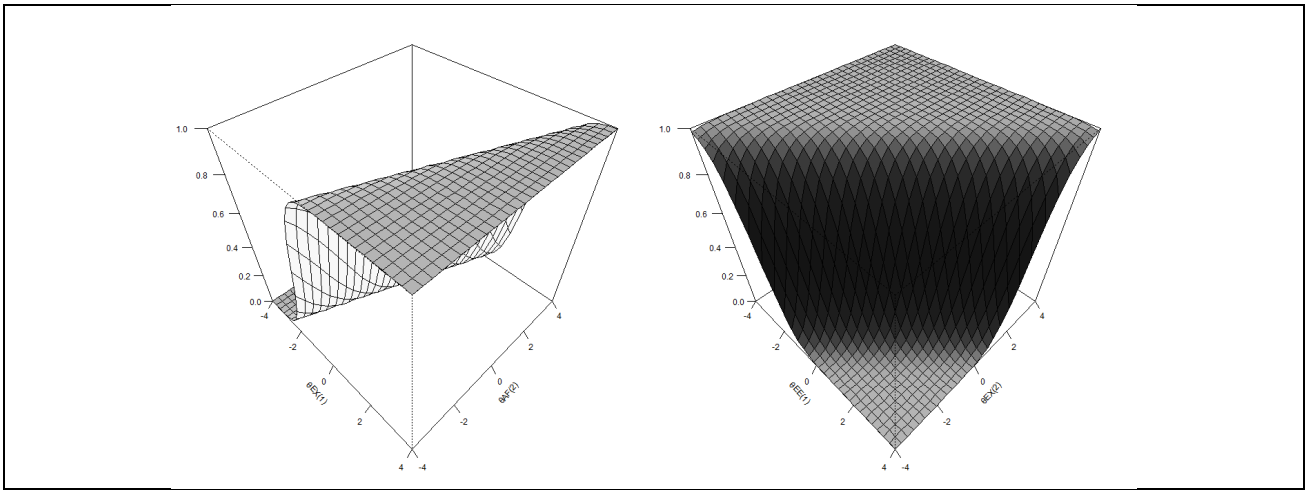
donde $\psi_{logístico}$ refleja la función logística y $\mathbf{a'}_2 - \mathbf{a'}_1$ es un vector de diferencias entre los parámetros *a* de los ítems del bloque. En el caso de que los ítems sean unidimensionales, el modelo se reduce a:

$$P(2[1,2]|\boldsymbol{\theta}) = \psi_{logistic}(\mathrm{a}_{d2}\theta_{d2} - \mathrm{a}_{d1}\theta_{d1} + c)$$

Por tanto, se establece que el *logit* de la probabilidad de elegir el ítem 2 del bloque se relaciona linealmente con los niveles de rasgo medidos por los ítems del bloque. El parámetro *c* determina la probabilidad de elección del ítem 2 cuando los niveles de rasgo son cero (la probabilidad sería expit(*c*); por ejemplo, si *c* = 0, la probabilidad es 0.5). La Figura 3 muestra la representación de la superficie de respuesta de dos bloques binarios según el MUPP.

Figura 3

Representación de la superficie de respuesta en el MUPP-2PL que representa la probabilidad de elegir el ítem 2 para un bloque con parámetros $a_{EX(1)}$ = -2.81, $a_{AF(2)}$ = -1.11 y c = -0.86 (izquierda) y para otro con parámetros $a_{EE(1)}$ = 2.21, $a_{EX(2)}$ = 2.25 y c = 4.30 (derecha).



En este caso, se representan las superficies de respuestas para los mismos bloques mostrados en la Figura 2, pero según la calibración del MUPP-2PL. En el primer bloque, la probabilidad de elegir el segundo ítem depende principalmente de la primera dimensión ($\theta_{EX(1)}$) y, en menor medida, de la segunda ($\theta_{AF(2)}$). Dado que ambos ítems son inversos (parámetros a negativos), la probabilidad de elegir el ítem 2 aumenta con $\theta_{EX(1)}$ y disminuye con $\theta_{AF(2)}$. En el segundo bloque, la probabilidad de elegir el segundo ítem depende en igual medida de ambas dimensiones ($\theta_{EE(1)}$ y $\theta_{EX(2)}$). Como los ítems son directos (parámetros a positivos), la probabilidad de elegir el ítem 2 aumenta con $\theta_{EX(2)}$ y disminuye con $\theta_{EE(1)}$. Además, se observa que, para niveles de rasgo iguales a cero, la probabilidad de elegir el segundo ítem es mayor en el segundo bloque que en el primero, lo que se explica por un mayor parámetro *c* en este bloque.

Como se observa al comparar las Figuras 2 y 3, ambos modelos, MUPP-2PL y TIRT, generan superficies de respuesta muy similares para comparaciones binarias y, en esencia, son prácticamente equivalentes (i.e., difieren únicamente en la función de enlace: *probit* o *logit* y en el procedimiento de estimación). Por lo tanto, en el caso de comparaciones binarias, la elección entre uno u otro modelo puede depender principalmente de las preferencias del investigador. Por ejemplo, una ventaja del TIRT es que sus parámetros están en una métrica fácilmente comprensible, ya que se asemejan a los pesos factoriales con los que los investigadores suelen estar familiarizados.

En escenarios más complejos, como comparaciones de más de dos elementos, las predicciones de ambos modelos pueden diferir. En el caso del TIRT, dichas comparaciones se traducen en comparaciones binarias, lo que introduce desafíos adicionales, como tratar la dependencia entre comparaciones (i.e., la preferencia por el ítem 1 frente al ítem 2 no es independiente de la preferencia por el ítem 1 frente al ítem 3) y la necesidad de imponer restricciones de igualdad en los parámetros de los ítems dentro de un mismo bloque (i.e., el peso del ítem 1 en un bloque debe ser igual en todas las comparaciones binarias que involucren dicho ítem). Por otro lado, el MUPP ofrece flexibilidad en la elección del modelo que siguen los ítems del bloque (e.g., punto ideal vs. dominancia). Sin embargo, esta flexibilidad viene acompañada de mayores desafíos en la estimación de los parámetros de los ítems (Zheng et al., 2024).

El uso de un modelo de TRI permite optimizar la estimación de los niveles de rasgo y además permite obtener medidas del error típico de estimación para cada nivel de rasgo (i. e., teniendo en cuenta que la precisión puede ser distinta en función de los bloques aplicados), así como medidas de fiabilidad marginal. Además, los modelos presentados anteriormente permiten el cálculo de los valores esperados de los parámetros de bloques a partir de los parámetros de ítems Likert. Esto ofrece la posibilidad de anticipar la capacidad informativa de los diferentes bloques posibles, con la finalidad de optimizar las pruebas de elección forzosa para una máxima precisión de cada evaluado (Kreitchmann et al., 2019, 2023).

Ensamblaje de bloques de elección forzosa y formato de respuesta

Aunque los modelos psicométricos de elección forzosa permitan capturar los matices en las relaciones entre rasgos e ítems, es el diseño del cuestionario lo que garantizará que estas relaciones sean efectivamente variadas, de forma que se reduzca la ipsatividad en las puntuaciones estimadas. Asimismo, el diseño de los bloques afectará a la robustez al falseamiento. El desarrollo de una prueba de elección forzosa optimizada requiere una serie de pasos: (a) crear un extenso banco de ítems calibrado mediante un modelo de TRI a partir de una muestra lo suficientemente amplia; (b) recopilar valoraciones de deseabilidad social de los ítems a través de un grupo de expertos; (c) utilizar un procedimiento de emparejamiento óptimo que respete las restricciones impuestas por las diferencias en deseabilidad social y, al mismo tiempo, maximice la fiabilidad; y (d) evaluar las propiedades psicométricas del test final en una muestra empírica. Con respecto al modo en que se conforman los bloques, hay ciertos aspectos a considerar:

Restricciones de deseabilidad social. Los ítems de los bloques deben emparejarse según su nivel de deseabilidad social. El primer paso consiste en obtener valoraciones de expertos sobre el grado de deseabilidad social de los enunciados (e.g., 1: altamente indeseable; 5: altamente deseable en un contexto específico (i.e., en

función de las características de un puesto de trabajo que puedan ser más deseables). Estas medidas pueden complementarse con correlaciones entre las valoraciones promedio de deseabilidad social de los expertos y la correlación entre las respuestas reales de los ítems con escalas tradicionales de deseabilidad social. En segundo lugar, es aconsejable utilizar medidas de similaridad de la deseabilidad social de los ítems del bloque, que tengan en cuenta no sólo la equiparación del puntaje medio en deseabilidad social de los ítems, sino también el consenso entre los jueces. Pavlov et al. (2021) y Pavlov (2024) proporcionan más detalles sobre el proceso de igualación en deseabilidad social, proponiendo el uso del índice linealmente ponderado de Brennan-Prediger (BPi; e.g., Brennan y Prediger, 1981; Gwet, 2014) como medida del acuerdo entre ítems en deseabilidad social. Este índice es similar al coeficiente kappa ponderado, pero la probabilidad esperada de acuerdo por azar se calcula asumiendo una distribución uniforme. Pavlov et al. (2021) recomiendan emparejar bloques por deseabilidad social usando un BPi significativamente mayor a 0.70 (con un intervalo de confianza del 95%, ya que el valor máximo, 1, indica un acuerdo perfecto, 0 indica ausencia de acuerdo y los valores negativos indican un desacuerdo sistemático).

Inclusión o no de bloques heteropolares (o mixtos). En la investigación previa, se distingue entre bloques homopolares positivos, homopolares negativos y heteropolares (mixtos). Los bloques homopolares están compuestos por ítems que miden diferentes dimensiones en la misma dirección. Por ejemplo, un bloque homopolar positivo podría estar formado por los ítems: "*Disfruto socializando con grupos grandes de personas*" (EX+) y "*Soy capaz de relajarme bajo presión*" (EE+), ambos midiendo su rasgo, extraversión y estabilidad emocional, respectivamente, de manera directa. Un bloque homopolar negativo estaría for-mado por ítems inversos, por ejemplo: "*Evito participar en conversa-ciones grupales*" (EX-) y "*Me resulta difícil mantener la calma en situaciones estresantes*" (EE-). Este tipo de blo-ques suele facilitar el emparejamiento de ítems, ya que ge-neralmente presentan una deseabilidad social similar. Los bloques heteropolares contendrían ítems que miden sus respectivas dimensiones en direcciones opuestas. Un ejemplo de bloque heteropolar, midiendo extraversión y responsabilidad, sería: "*Prefiero evitar conversaciones con personas que no conozco*" (EX-) y "*Me considero alguien con mucha determinación*" (RE +). De acuerdo con Brown y Maydeu-Olivares (2011), la inclusión de bloques heteropolares es necesaria para identificar con mayor precisión la posición absoluta de una persona en el continuo de rasgos (en la Tabla 1, vimos un ejemplo de uso de bloques exclusivamente homopolares). Sin embargo, lograr una deseabilidad social equilibrada en ítems opuestos es un reto, ya que los participantes pueden identificar el ítem más deseable y sesgar sus respuestas, comprometiendo así la validez de la prueba. Llegados a este punto, parecería que nos encontramos en una encrucijada: tener que elegir entre una prueba más falseable (con bloques heteropolares) o una con puntuaciones más ipsativas (formada exclusivamente de bloques homopolares). Li et al. (2024) sugieren que es posible alcanzar un equilibrio entre la resistencia al falseamiento y las propiedades psicométricas, ya que el uso de un 20% de bloques heteropolares es suficiente para optimizar la precisión en la medición, manteniendo a la vez la resistencia al falseamiento mediante una alta coincidencia en la deseabilidad social entre los ítems del mismo bloque. Por otro lado, algunos autores han mostrado que un emparejamiento óptimo de bloques homopolares puede ser suficiente para reducir el problema de la ipsatividad, tanto mediante estudios de simulación (Kreitchmann et al., 2022; 2023), como con datos empíricos (Graña et al., 2024). En este sentido, resulta clave que se escojan ítems que miden las dimensiones en distintas magnitudes, que en la práctica puede simplificarse a emparejar ítems con mayor diferencia de pesos intra-bloque (i.e., emparejar un ítem que pesa alto en una dimensión, con otro que pesa bajo en la otra dimensión; y viceversa).

Algoritmos de emparejamiento. El número de bloques conformables a partir de un conjunto amplio de ítems es muy elevado. Por ejemplo, ensamblar 60 ítems en 30 blo-ques de 2 deriva, aproximadamente, en 2,92 X 10^{40} cues-tionarios posibles (Kreitchmann et al., 2021). Por lo tanto, se hace necesario un algoritmo para emparejarlos. Más allá de la inclusión o no de bloques heteropolares, es espe-cialmente relevante el uso de algoritmos que optimicen el emparejamiento de bloques. Li et al. (2024) presentan un tutorial que explica cómo construir estos cuestionarios y evaluar su calidad a través de simulaciones, utilizando el paquete de R *autoFC* (Li et al., 2022).

Por su parte, Kreitchmann et al. (2021) adaptaron un algoritmo genético NHBSA (algoritmo de muestreo basado en histogramas de nodos; Tsutsui, 2006) al desafío de ensamblar ítems en bloques, proporcionando una implementación accesible a través de Shiny, con el fin de facilitar el diseño de pruebas de elección forzosa (https://psychometricmodelling.shinyapps.io/FCoptimization/). Estos procesos de ensamblaje también pueden utilizarse para generar bancos de bloques óptimos, que dispongan de un número amplio de bloques elegibles para cada nivel de rasgo, y que pue-dan servir de base para test adaptativos informatizados (Kreitchmann et al., 2023).

Formato de respuesta y estructura de los bloques. Otra decisión por tomar es el tamaño de los bloques; esto es, se pueden constituir pares, tripletas o tétradas. El uso de tripletas puede reducir la ipsatividad, pero algunos autores señalan que incrementa la carga cognitiva de las personas evaluadas al requerir más comparaciones por bloque (Sass et al., 2020); además el uso de tripletas puede dar lugar a modelos más complejos, incluso si se utiliza el TIRT, ya que este ignora las correlaciones entre pares en la estimación del nivel de rasgo.

Por otro lado, el formato de elección binaria puede dar lugar a una disminución de la fiabilidad de las puntuaciones, lo que puede esperarse como consecuencia inherente a la naturaleza dicotómica de las respuestas. Como alternativa, Brown y Maydeu-Olivares (2018b) proponen el uso de un formato de respuesta graduada, en el que los participantes expresan sus preferencias utilizando varias categorías. Este formato ya ha mostrado algunos resultados prometedores (Zhang et al., 2024), ya que combina las ventajas de los bloques de elección forzosa (mejor control de la deseabilidad social) y de las escalas tipo Likert (mayor número de categorías para diferenciar mejor las respuestas). Zhang et al. (2024) ofrecen un análisis exhaustivo de este formato, demostrando su utilidad en mejorar la precisión y la validez de las pruebas en comparación con los formatos tradicionales de elección forzosa y de Likert.

Otros factores. Otros factores que influyen en la ipsatividad incluyen las correlaciones entre las dimensiones evaluadas y la cantidad de dimensiones que se miden. A medida que disminuye el número de dimensiones o aumentan las correlaciones positivas entre ellas, la ipsatividad tiende a incrementarse. Esto ocurre porque una menor diferenciación entre las dimensiones reduce la capacidad de capturar de manera independiente cada rasgo. Por ejemplo, en sus estudios de simulación, Bürkner et al. (2019) concluyen que, con cinco o menos dimensiones y utilizando bloques homopolares, es difícil obtener mediciones precisas. Sin embargo, cuando se miden 30 dimensiones, los resultados mejoran significativamente, logrando una buena recuperación de los rasgos evaluados. Cabe destacar que estos autores no analizaron casos intermedios, es decir, aquellos con entre 6 y 29 dimensiones, lo que deja abierta la posibilidad de estudios adicionales para explorar el comportamiento en ese rango.

Calibración de los modelos

La calibración de estos modelos puede ser más o menos compleja, en función del diseño y modelos elegidos. Si se utiliza el TIRT, con un diseño de tripletas, deben elegirse algunas restricciones para la identificación del modelo. De acuerdo con Jansen y Schulze (2024) estas restricciones pueden ser problemáticas a la hora de recuperar los parámetros. En cuanto al software, Brown y Maydeu-Olivares (2012) proporcionan una macro en Excel (http://annabrown.name/software) que, una vez introducida la información básica sobre el diseño de bloques, genera una sintaxis de Mplus (Muthén y Muthén, 2018). Igualmente, puede utilizarse el paquete *lavaan* (Rosseel, 2012), puesto que el TIRT puede entenderse como un modelo de ecuaciones estructurales para variables categóricas (estimable, por ejemplo, con el método Mínimos Cuadrados Ponderados Robustos [WLSMV], sobre las correla-ciones policóricas).

Algunos autores han encontrado problemas en estas aproximaciones, principalmente las bajas tasas de convergencia de los modelos, el coste computacional y una cierta dependencia de los valores fijados, lo que se considera problemático (ver Bürkner et al., 2019). Otra aproximación posible es el uso del algoritmo de Monte Carlo basado en cadenas de Markov (MCMC), aunque con el problema asociado del alto coste computacional. Nie et al. (2024) proporcionan una tabla de resumen de distintas aproximaciones para la calibración. Otra posibilidad consiste en utilizar un pro-

cedimiento en dos fases. En primer lugar, se precalibran las respuestas a los ítems Likert; estas estimaciones se integran posteriormente en la estructura del modelo, dándolos como fijas. Esa aproximación asume invarianza de los parámetros, lo que puede no ser correcto. No obstante, algunos estudios han mostrado niveles de invarianza elevados (ver Morillo et al., 2019).

Estimación del nivel de rasgo y de la fiabilidad de la prueba

Para estimar los niveles de rasgos latentes se pueden emplear los métodos de máxima verosimilitud (MLE), a posteriori máximo (MAP) y a posteriori esperado (EAP). En pruebas con un alto número de dimensiones, es recomendable optar por MAP o MLE, ya que el coste computacional de EAP aumenta significativamente. Esto se debe a que el número de puntos de cuadratura necesarios crece exponencialmente con el número de dimensiones (e.g., con 25 puntos de cuadratura por dimensión [D], el total de puntos de cuadratura asciende a 25^D).

Invarianza de los parámetros

La comprobación de la invarianza en los parámetros de los ítems al ensamblarse en bloques es esencial para lograr un ensamblaje exitoso y hacer viable la creación de pruebas de elección forzosa *on-the-fly* en contextos de selección de personal. En este caso, hablamos de un proceso de ensamblaje en tiempo real, permitiendo seleccionar dinámicamente el bloque más adecuado para cada evaluado según sus respuestas anteriores. Este proceso de ensamblaje en tiempo real optimizaría tanto la precisión como la eficiencia del cuestionario.

La invarianza es importante porque asegura que las propiedades psicométricas de los bloques sean consistentes y no dependan de las combinaciones específicas de ítems en las que se presentan. Este aspecto puede analizarse mediante estrategias tradicionales de funcionamiento diferencial de los ítems, utilizando modelos de TRI. Por ejemplo, Morillo et al. (2019) emplean el test de razón de verosimilitudes, comparando un modelo restringido (donde todos los parámetros son idénticos entre las versiones Likert y de elección forzosa) con modelos en los que un parámetro particular varía entre versiones. Lin y Brown (2017), por su parte, estiman los parámetros de los ítems en bloques de tres y los comparan con los obtenidos en bloques de cuatro, tras realizar una equiparación de parámetros para asegurar su comparabilidad. Estos estudios encontraron cierto grado de invarianza en bloques binarios (Morillo et al., 2019), aunque también evidencian que los parámetros pueden variar en función del contexto de los bloques (Lin y Brown, 2017). La invarianza tiende a ser más estable para los parámetros de discriminación que para los de umbral (Lin y Brown, 2017). Estos resultados sugieren que el problema puede mitigarse mediante un diseño cuidadoso de los bloques, evitando combinaciones de ítems que puedan inducir comparaciones no deseadas. Es esperable que el problema del contexto sea mayor para bloques no binarios. En cualquier caso, una solución sería implementar correcciones a través de modelos TRI que consideren las variaciones de parámetros derivadas del contexto.

Pasos para la construcción de pruebas de elección forzosa

A partir de los aspectos relevantes definidos anteriormente, podemos plantear una serie de pasos recomendados a seguir a la hora de construir una prueba de elección forzosa.

Construcción y calibración del banco de ítems. Con la finalidad de poder posteriormente diseñar la prueba de elección forzosa en base a sus propiedades psicométricas, se recomienda partir de un banco de ítems administrados individualmente en formato de respuesta graduada (grado de acuerdo) sobre una muestra en la que se minimicen ses-gos como la deseabilidad social. A través de valoraciones de expertos, se obtiene información acerca de la deseabi-lidad social de los ítems para el contexto en el que se busca utilizar la prueba, que servirá para equiparar la deseabili-dad de los ítems que formarán parte de un mismo bloque.

Definición de especificaciones de la prueba. Respetando las limitaciones del banco de ítems, se determina el número de bloques deseado, la representatividad de las

dimensiones en ellos, y las restricciones de deseabilidad social. La elección del modelo depende del tipo de ítems (dominancia o punto ideal) y del formato de respuesta (pares o triadas). Por ejemplo, el modelo TIRT admite bloques con más de dos ítems, mientras que el MUPP-2PL se limita a pares; por otro lado, el MUPP permite la utilización de ítems de punto ideal, mientras el TIRT se limita a ítems de dominancia.

Ensamblaje de la prueba. Para el ensamblaje óptimo de un test, cabe utilizar la información acerca de las propiedades psicométricas de los ítems en la calibración del banco para predecir las propiedades psicométricas (discriminación, umbral, información) de los bloques construidos. Así, es posible generar una prueba que maximice la fiabilidad de las puntuaciones, a la vez que cumple con restricciones de deseabilidad social y de contenido de los bloques definidas en la fase de especificación. En esta fase, cabe considerar también la posibilidad de la aplicación adaptativa de la prueba, de forma que sea posible administrar a cada evaluado aquellos bloques que permitan estimar su puntuación con la mayor precisión. El ensamblaje de los bloques en una administración adaptativa puede hacerse en directo (*on-the-fly*), considerando todas las posibles combinaciones de ítems que cumplan con criterios de equiparación en deseabilidad social.

Administración y calibración de la prueba. Los datos recogidos de las respuestas de elección forzosa deben utilizarse para calibrar los parámetros en este formato, buscando identificar posibles dependencias contextuales de los ítems. Por ejemplo, la discriminación o el parámetro de umbral de un ítem puedan variar según con qué ítem éste se empareje.

Cálculo de puntuaciones y evaluación de la calidad métrica. Con los datos de la prueba, se estiman las puntuaciones de los sujetos, evaluando su fiabilidad, así como el ajuste del modelo y la invarianza de los parámetros. En pruebas adaptativas, la invarianza es crucial para garantizar que las puntuaciones estimadas se basen en parámetros consistentes.

Ilustración empírica

Con la finalidad de ilustrar los principales pasos en la construcción de una prueba de elección forzosa, se incluyen datos de ejemplos y códigos de R a través de https://osf.io/a5tsg/. Asimismo, se incluye un tutorial comentado sobre cada apartado del código, con resultados e interpretaciones.

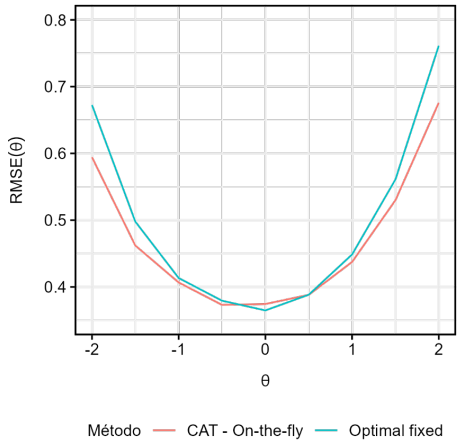
En el ejemplo, se parte de una selección de los datos del IPIP-NEO recogidos por Johnson (2014). El conjunto original de datos incluye 300 ítems del IPIP-NEO y 307313 participantes (https://osf.io/tbmh5). En nuestro caso, se seleccionaron los datos de una muestra aleatoria de 1000 participantes sin valores perdidos, recogidos en Estados Unidos y de entre 19 y 24 años. Se seleccionaron los 20 ítems que más claramente pesaban en su dimensión teórica (de acuerdo con el análisis factorial en una muestra aleatoria distinta, de 30000 participantes). Los datos de deseabilidad social de los ítems del IPIP-NEO se han to-mado del trabajo de Hughes et al. (2021; https://osf.io/8gfxs/).

Así, se ha logrado un banco de 100 ítems midiendo los Cinco Grandes (EE: estabilidad emocional, EX: extraversión; AP: apertura a la experiencia; AF: afabilidad; RE: responsabilidad), con pesos factoriales con media de 0.66 y desviación típica de 0.11, ajuste próximo a lo aceptable (RMSEA = 0.062; CFI = 0.891), y alta fiabilidad de las puntuaciones en todas las dimensiones, entre 0.90 para *apertura* y 0.95 para *extroversión*.

A partir del banco, se ha construido una prueba de 40 bloques binarios utilizando el modelo MUPP y el algoritmo de ensamblaje disponible en https://psychometricmodelling.shinyapps.io/blockAssemblySD/. A través de un estudio simulación se encuentra una fiabilidad empírica esperada media-alta con el cuestionario generado (EE: 0.81; EX: 0.83; AP: 0.74; AF: 0.71; RE: 0.78). Se ilustra también la utilización de un *test* adaptativo *on-the-fly*, con el que se obtienen fiabilidades mayores que para un test de elección forzosa fijo (EE: 0.84; EX: 0.86; AP: 0.75; AF: 0.74; y RE: 0.81), y especialmente mejores estimaciones para niveles de rasgo más alejados del promedio (Figura 4).

Figura 4

Promedio del RMSE (raíz del error cuadrático medio) para cada nivel de rasgo (valor promedio a través de los rasgos) y para cada tipo de test (fijo óptimo o adaptativo on-the-fly)



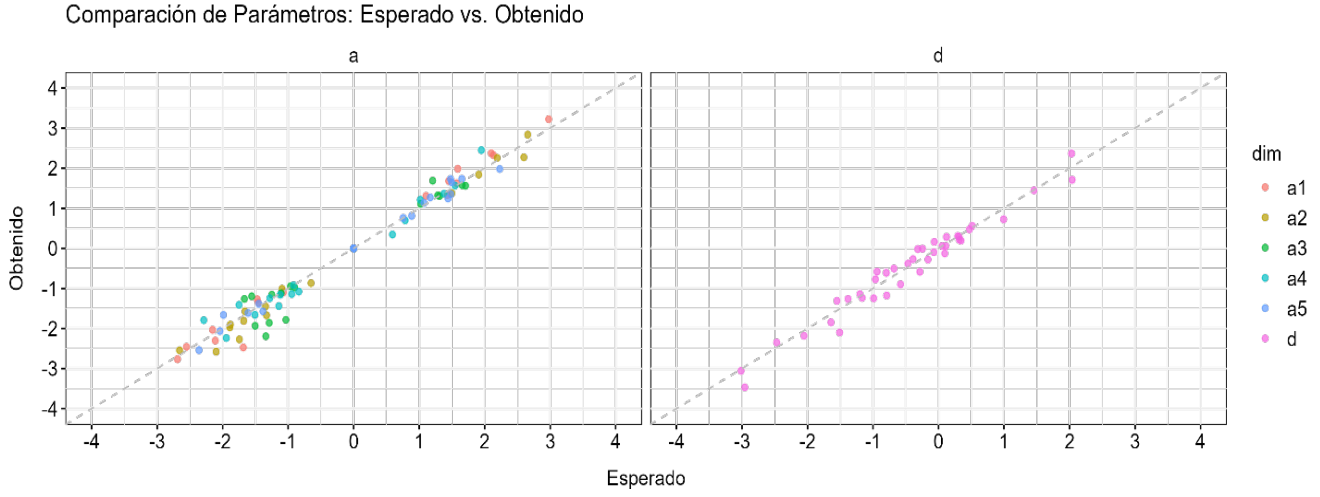
Por último, en la ilustración, con datos simulados, se observa el cumplimiento de la invarianza de los parámetros, ya que los parámetros de discriminación (a) y umbral (d) estimados con los datos de respuestas de elección forzosa se aproximan a su valor esperado a partir de la generalización de los parámetros de los ítems bajo el modelo MUPP-2PL (Figura 5).

Discusión

Los modelos recientes de TRI para elección forzosa, como el TIRT y el MUPP, han mejorado la precisión en la estimación de los rasgos evaluados. Además, estos modelos permiten el ensamblaje óptimo de pruebas mediante

Figura 5

Comparación de parámetros esperados y obtenidos en los bloques



algoritmos de emparejamiento de ítems, en los que se controla la similitud en deseabilidad social de los ítems emparejados, al tiempo que se maximiza la fiabilidad de la prueba completa. Nuestra perspectiva es que, si bien el uso de bloques homopolares y heteropolares puede reducir considerablemente la ipsatividad de las puntuaciones, la implementación de algoritmos de optimización podría alcanzar el mismo objetivo empleando únicamente bloques homopolares, más robustos a la deseabilidad social. Además, el uso de la TRI ofrece una ventaja adicional: facilita la construcción de test adaptativos informatizados y la implementación de pruebas adaptativas *on-the-fly*, que ajustan la prueba en tiempo real y de forma dinámica para cada evaluado, maximizando la precisión de las puntuaciones. Sin embargo, el beneficio de estos sistemas avanzados depende en gran medida de la disponibilidad de un banco de ítems amplio (e.g., con variabilidad en los parámetros de los ítems) y de consideraciones sobre el coste computacional, que puede incrementarse considerablemente en bancos de gran tamaño. Paralelamente, el emparejamiento adecuado de ítems sigue siendo un aspecto crítico para su efectividad. La construcción de estas pruebas es más exigente, tanto por las consideraciones técnicas mencionadas como por el hecho de que, en formatos de bloques de opción binaria, la información por unidad tiende a ser menor, lo cual suele requerir pruebas de mayor longitud. Finalmente, el éxito del ensamblaje y la construcción de prue-bas adaptativas *on-the-fly* dependerá también del supuesto de invarianza de parámetros en distintos contextos de aplicación, un aspecto que aún requiere mayor exploración en futuras investigaciones.

En este artículo, además de discutir los diferentes aspectos relevantes en la evaluación con pruebas de elección forzosa, se presentan recomendaciones de los pasos a seguir que han mostrado ser eficaces para la construcción de pruebas de elección forzosa de personalidad (Graña et al., 2024). Además, a través de la accesibilidad de código abierto en R, se busca facilitar a que profesionales aplica-dos puedan seguir con facilidad estas recomendaciones para así construir pruebas robustas a los sesgos de respues-tas, con máxima fiabilidad y validez.

Referencias

Abad, F. J., Kreitchmann, R. S., Sorrel, M. A., Nájera, P., García-Garzón, E., Garrido, L. E. y Jiménez, M. (2022). 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]. *Papeles del Psicólogo, 43*(1), 29–35. https://doi.org/10.23923/pap.psicol.2982

Brennan, R. L. y Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement, 41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Brown, A. y 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. y 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. y 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. y Maydeu-Olivares, A. (2018a). Modelling forced-choice response formats. En P. Irwing, T. Booth y D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 523–569). Wiley. https://doi.org/10.1002/9781118489772.ch18

Brown, A. y Maydeu-Olivares, A. (2018b). Ordinal Factor Analysis of Graded-Preference Questionnaire Data. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247

Bürkner, P.-C., Schulte, N. y 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. y 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

Graña, D. F., Kreitchmann, R. S., Abad, F. J. y Sorrel, M. A. (2024). Equally vs. Unequally Keyed Blocks in Forced-Choice Questionnaires: Implications on Validity and Reliability. *Journal of Personality Assessment,* 1–14. https://doi.org/10.1080/00223891.2024.2420869

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Heggestad, E. D., Morrison, M., Reeve, C. L. y 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. y 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. y 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

Hughes, A. W., Dunlop, P. D., Holtrop, D. y Wee, S. (2021). Spotting the "ideal" Personality Response: Effects of Item Matching in Forced Choice Measures for Personnel Selection. *Journal of Personnel Psychology, 20*(1), 17–26. https://doi.org/10.1027/1866-5888/a000267

Jansen, M. T. y Schulze, R. (2023). Linear Factor Analytic Thurstonian Forced-Choice Models: Current Status and Issues. *Educational and Psychological Measurement*, *84*(4), 660–690. https://doi.org/10.1177/00131644231205011

Johnson, J. A. (2014). Measuring Thirty Facets of the Five Factor Model with a 120-item public Domain Inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Kreitchmann, R. S., Abad, F. J. y Sorrel, M. A. (2022). A Genetic Algorithm for Optimal Assembly of Pairwise Forced-Choice Questionnaires. *Behavior Research Methods, 54,* 1476–1492. https://doi.org/10.3758/s13428-021-01677-4

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D. y Morillo, D. (2019). Controlling for Response Biases in Self-Report Scales: Forced-choice vs. Psychometric Modeling of Likert Items. *Frontiers in Psychology, 10,* Artículo 2309. https://doi.org/10.3389/fpsyg.2019.02309

Kreitchmann, R. S., Sorrel, M. A. y Abad, F. J. (2023). On Bank Assembly and Block Selection in Multidimensional Forced-Choice Adaptive Assessments. *Educational and Psychological Measurement, 83*(2), 294–321. https://doi.org/10.1177/00131644221087986

Li, M., Sun, T. y Zhang, B. (2022). autoFC: An R Package for Automatic Item Pairing in Forced-Choice Test Construction. *Applied Psychological Measurement, 46*(1), 70–72. https://doi.org/10.1177/01466216211051726

Li, M., Zhang, B., Li, L., Sun, T. y Brown, A. (2024). Mix-keying or Desirability-Matching in the Construction of Forced-Choice Measures? An Empirical Investigation and Practical Recommendations. *Organizational Research Methods, 0*(0). Advance Online Publication. https://doi.org/10.1177/10944281241229784

Lin, Y. 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

Martínez, A. y Salgado, J. F. (2021). A Meta-Analysis of the Faking Resistance of Forced-Choice Personality Inventories. *Frontiers in Psychology, 12,* Artículo 732241. https://doi.org/10.3389/fpsyg.2021.732241

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P. y 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. y 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

Muthen, L. K. y Muthen, B. O. (2018). *Mplus User's Guide* (8ª ed.). Muthen & Muthen.

Nie, L., Xu, P. y Hu, D. (2024). Multidimensional IRT for Forced Choice Tests: A Literature Review. *Heliyon, 10*(5), Artículo e26884. https://doi.org/10.1016/j.heliyon.2024.e26884

Pavlov, G. (2024). An Investigation of Effects of Instruction Set on Item Desirability Matching. *Personality and Individual Differences, 216,* Artículo 112423. https://doi.org/10.1016/j.paid.2023.1124233

Pavlov, G., Shi, D., Maydeu-Olivares, A. y Fairchild, A. (2021). Item Desirability Matching in Forced-Choice Test Construction. *Personality and Individual Differences, 183,* Artículo 111114. https://doi.org/10.1016/j.paid.2021.111114

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sass, R., Frick, S., Reips, U.-D. y 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. y 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

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.2006.1688544

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63*(2), 117–124. https://doi.org/10.1037/h0021567

Zhang, B., Luo, J. y Li, J. (2024). Moving Beyond Likert and Traditional Forced-Choice Scales: A Comprehensive Investigation of the Graded Forced-Choice Format. *Multivariate Behavioral Research, 59*(3), 434–460. https://doi.org/10.1080/00273171.2023.2235682

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S. y White, L. A. (2020). Though Forced, Still Valid: Psychometric Equivalence of Forced-Choice and Single-Statement Measures. *Organizational Research Methods, 23*(3), 569–590. https://doi.org/10.1177/1094428119836486

Zheng, C., Liu, J., Li, Y., Xu, P., Zhang, B., Wei, R., Zhang, W., Liu, B. y Huang, J. (2024). A 2PLM-RANK Multidimensional Forced-Choice Model and its Fast Estimation Algorithm. *Behavior Research Methods, 56,* 6363–6388. https://doi.org/10.3758/s13428-023-02315-x